2019/6/6　北大SPH・統計解析の応用⑤⑥

観察研究に生じるバイアスと制御

北海道大学　医学統計学
横田　勲

今回の内容のポイント

2

- 因果推論の入門的話題をとして、以下の2つの因果モデルを扱う
  - 反事実アウトカムモデル、潜在アウトカム
  - 因果ダイアグラム
- 標的集団
  - どのような集団における効果なのか？
- 想定する因果関係の図示

Hさんがジムに！

4

- Hさんはなんと7kgもやせた！（事実）
  - さすがはプライベートジム！
- ジムは効果がある
  - 他の人にも同様・・？
- ジムはHさんには効果がある
  - ただの正月太りが戻っただけかも・・？

どちらも結論できない

コントロール control

5

- 比較する相手
  - 対照
- 何かの効果を調べる場合には、比較が基本

Hさん個人について

6

- ジムでやせるか？
- 2019年1月1日

ジムに通う
Hさん

ジムに通わない
Hさん

2019年5月1日
やせたかどうかを調べる

反事実アウトカム

7

- 事実が観察されたら、観察されないアウトカム
- 事実データ factual data
  - Hさんがジムに通って（$a_H = 1$）、4ヶ月後にやせた（$Y_H = 1$）
- 反事実データ counterfactual data
  - Hさんがジムに通わなかったら（$a_H = 0$）、4ヶ月後には？（$Y_H = ?$）

潜在アウトカム potential outcome

8

- $Y^{a=1}$
  - 曝露$a=1$を受けた場合のアウトカム
- $Y^{a=0}$
  - 曝露$a=0$を受けた場合のアウトカム
- アウトカムも2値(0,1)の場合

	$Y^{a=1}$	$Y^{a=0}$
Doomed	1	1
Helped	1	0
Hurt	0	1
Immune	0	0

潜在アウトカムと観察アウトカム

9

- 受けた曝露に応じて、
潜在アウトカムのいずれかが観察される

	A	$Y^{a=1}$	$Y^{a=0}$	Y
Doomed	1	1	1	1
Helped	1	1	0	1
Hurt	1	0	1	0
Immune	1	0	0	0
Doomed	0	1	1	1
Helped	0	1	0	0
Hurt	0	0	1	1
Immune	0	0	0	0

個人での因果効果

10

	$Y^{a=1}$	$Y^{a=0}$	Causal effect $Y^{a=1}-Y^{a=0}$
Doomed	1	1	$1-1=0$
Helped	1	0	$1-0=1$
Hurt	0	1	$0-1=-1$
Immune	0	0	$0-0=0$

- データとして観察はできない
  - 反事実アウトカムとの比較で定義可能
- Sharp causal null hypothesis
  - Doomed, Immuneな人しかいない

平均因果効果 Average Causal Effects

11

- $E[Y^{a=1}]-E[Y^{a=0}]$
  - 集団全員が曝露を受けた場合と
集団全員が曝露を受けなかった場合の差
- Null hypothesis of no average causal effect
  - $E[Y^{a=1}]=E[Y^{a=0}]$
  - Sharp causal null hypothesisに加え、
Helpedな人とHurtな人が同数いる場合も成立

因果効果の指標

12

- 因果リスク差
  - $\Pr[Y^{a=1}=1]-\Pr[Y^{a=0}=1]$
- 因果リスク比
  - $\dfrac{\Pr[Y^{a=1}=1]}{\Pr[Y^{a=0}=1]}$
- 因果オッズ比
  - $\dfrac{\Pr[Y^{a=1}=1]/\Pr[Y^{a=1}=0]}{\Pr[Y^{a=0}=1]/\Pr[Y^{a=0}=0]}$

練習① 以下のACEは？

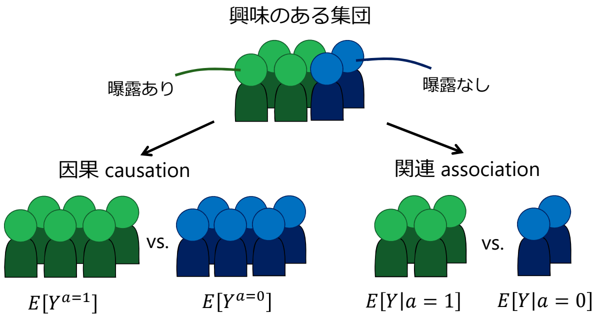
13

ID	$Y^{a=1}$	$Y^{a=0}$
1	0	1
2	1	0
3	0	0
4	0	0
5	0	0
6	1	0
7	0	0
8	0	1
9	1	1
10	1	0

ID	$Y^{a=1}$	$Y^{a=0}$
11	0	1
12	1	1
13	1	1
14	0	1
15	0	1
16	0	1
17	1	1
18	1	0
19	1	0
20	1	0

Association is not causation

14



Hernán MA, Robins JM (2019). Causal Inference. Chapman & Hall/CRC, forthcoming. を基に作成

関連効果の指標

15

- 関連リスク差
  - $\Pr[Y=1|A=1] - \Pr[Y=1|A=0]$
- 関連リスク比
  - $\dfrac{\Pr[Y=1|A=1]}{\Pr[Y=1|A=0]}$
- 関連オッズ比
  - $\dfrac{\Pr[Y=1|A=1]/\Pr[Y=0|A=1]}{\Pr[Y=1|A=0]/\Pr[Y=0|A=0]}$

因果効果指標と関連効果指標

16

- 因果効果指標は定義、概念的なもの
  - 反事実アウトカムを用いて定義されるため
- 関連効果指標は観察データから求まる
- 関連効果指標をもって
  因果効果指標を求めるには？
  - どのような条件が成立すれば？
  - どのような解析を行えば？

交絡 confounding

17

- 実際の曝露群での結果と
  集団全体が曝露した場合が違う
  - $E[Y^{a=1}|A=1] \neq E[Y^{a=1}]$

and / or

- 実際の非曝露群での結果と
  集団全体が曝露しなかった場合が違う
  - $E[Y^{a=0}|A=0] \neq E[Y^{a=0}]$

ランダム化による交換可能性の成立
exchangeability

18

- 曝露群での結果と
  非曝露群が、仮に曝露を受けた場合の結果
  が一致（その逆も）
  - $\Pr[Y^{a=1}|A=1] = \Pr[Y^{a=1}|A=0]$
    $\Pr[Y^{a=0}|A=0] = \Pr[Y^{a=0}|A=1]$
  - $Y^{a} \perp\!\!\!\perp A$ for all a
- 片方の集団と全体集団での結果と一致
  - $\Pr[Y^{a=1}|A=1] = \Pr[Y^{a=1}|A=0] = \Pr[Y^{a=1}]$
    $\Pr[Y^{a=0}|A=0] = \Pr[Y^{a=0}|A=1] = \Pr[Y^{a=0}]$

交換可能性の意味

19

- 実際の曝露と反事実アウトカムが独立
  - 曝露とアウトカムの関連ナシではない！
- 交絡が生じる状況では交換可能性が不成立
  - 曝露群には実はdoomedな人だらけ
    非曝露群には実はimmuneな人だらけ

条件付き交換可能性
20

- 予後因子Lが同じ値を持つ集団（層内）では交換可能性が成立
  - $\Pr[Y^{a=1}|A=1,L=1]=\Pr[Y^{a=1}|A=0,L=1]$
    $\Pr[Y^{a=0}|A=0,L=1]=\Pr[Y^{a=0}|A=1,L=1]$
  - $\Pr[Y^{a=1}|A=1,L=0]=\Pr[Y^{a=1}|A=0,L=0]$
    $\Pr[Y^{a=0}|A=0,L=0]=\Pr[Y^{a=0}|A=1,L=0]$
  - $Y^a \perp\!\!\!\perp A|L$ for all a
- No unmeasured confounding
  - 残差交絡 residual confounding がない

標準化 standardization
21

- 層ごとの結果の重み付き平均

$$\Pr[Y^a=1]=\sum_l \Pr[Y^a=1|L=l]\Pr[L=l]$$

  - 層別解析 stratified analysisのひとつ

練習②　標準化リスク差／比は？
22

ID	L	A	Y
1	0	0	0
2	0	0	1
3	0	0	0
4	0	0	0
5	0	1	0
6	0	1	0
7	0	1	0
8	0	1	1
9	1	0	1
10	1	0	1

ID	L	A	Y
11	1	0	0
12	1	1	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	1	1
17	1	1	1
18	1	1	0
19	1	1	0
20	1	1	0

回帰モデルによる標準化
23

- 層別しきれないほどの予後因子 $L_1, L_2, \cdots$
- $E(Y|A,L_1,L_2,\cdots)$を回帰モデルで表現
  - 例えば、一般化線形モデルの利用
- 個人iごとに反事実リスクを予測
  - $A=1$だった場合：
    $\hat{R}_i^{a=1}=E(Y|A=1,L_1=l_{1i},L_2=l_{2i},\cdots)$
  - $A=0$だった場合：
    $\hat{R}_i^{a=0}=E(Y|A=0,L_1=l_{1i},L_2=l_{2i},\cdots)$
- $\hat{R}_i^{a=1}-\hat{R}_i^{a=0}$を全員で平均

回帰モデルによる効果推定
24

- 例：ロジスティック回帰モデル

$$\log(\text{オッズ})=\beta_0+\beta_1\times A+\beta_2\times L$$

オッズ	$L=0$	$L=1$
$A=1$	$\exp(\beta_0+\beta_1)$	$\exp(\beta_0+\beta_1+\beta_2)$
$A=0$	$\exp(\beta_0)$	$\exp(\beta_0+\beta_2)$

$L=0$でのオッズ比

$$\frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)}=\mathbf{exp(\boldsymbol{\beta}_1)}$$

$L=1$でのオッズ比

$$\frac{\exp(\beta_0+\beta_1+\beta_2)}{\exp(\beta_0+\beta_2)}=\mathbf{exp(\boldsymbol{\beta}_1)}$$

モデルによる標準化との違い
25

- 回帰係数；条件付き因果効果
  - 交絡変数の水準が同じ部分集団での効果
  - 部分集団間で効果が等しいと仮定
    - 交互作用項を含めない場合
- 標準化；周辺因果効果
  - 集団全体が$A=1$だった場合と集団全体が$A=0$だった場合の差
  - 交絡変数の分布について周辺をとった

練習②スライドデータの要約 26

- $L=0$である8人の曝露状況
  - 曝露なし（$A=0$）が4人、イベント率25%
  - 曝露あり（$A=1$）が4人、イベント率25%
- $L=1$である12人の曝露状況
  - 曝露なし（$A=0$）が3人、イベント率67%
  - 曝露あり（$A=1$）が9人、イベント率67%

$L=1$である12人 27

- 12人全員が非曝露であったら、
  何名がイベントを起こすだろうか？
  - 実際の非曝露群では3人中2人
  - 交換可能性成立より、同じ割合でイベント発生
  - ゆえ、12人中8人がイベントを起こすだろう
- 12人全員が曝露であったら
  何名がイベントを起こすだろうか？
  - 実際の曝露群では9人中6人
  - 交換可能性より、12人中8人だろう

擬似集団 pseudo-population 28

- 因果リスクを知る上で必要な、
  全員が曝露／非曝露である場合の仮想集団
- $L=1$では、実際に曝露を受けた人は9/12
  - 曝露を受けた割合の逆数をかけてみよう
  - $\frac{1}{\frac{9}{12}}=\frac{12}{9}$ 倍
- 曝露を受けなかった人は3/12
  - 逆数をかけよう；$\frac{1}{\frac{3}{12}}=4$ 倍

逆確率重み付け法 29

- Inverse probability weighting method
  - IPW法
- 生成した擬似集団での関連効果の指標は
  - 擬似集団での因果効果の指標
  - 元の集団での因果効果の指標　に同じ
    - $Y^a \perp\!\!\!\perp A|L$ for all a　ゆえ

傾向スコア propensity score; PS 30

- 各層の曝露を受ける確率
- 個人単位で傾向スコアを予測するモデル
  を作ってもいい
  - L が連続量の場合、適当なカットオフ値で
    水準に分割する必要がなくなる

練習③　擬似集団を作ってみよう 31

ID	L	A	Y
1	0	0	0
2	0	0	1
3	0	0	0
4	0	0	0
5	0	1	0
6	0	1	0
7	0	1	0
8	0	1	1
9	1	0	1
10	1	0	1

ID	L	A	Y
11	1	0	0
12	1	1	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	1	1
17	1	1	1
18	1	1	0
19	1	1	0
20	1	1	0

PSの使い方

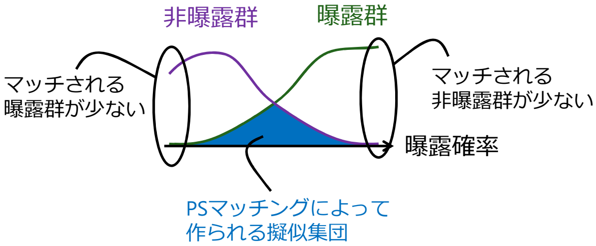
33

- 個人ごとに求まるため柔軟な使い方
- IPW解析
- マッチング
  - 近いPSをもつ曝露群と非曝露群の人々をペアに
- 層別解析
  - *L*が少ない場合、*L*での層別と等価
  - 傾向スコアが連続的に変化するなら、残差交絡（層内で均質でない）
- 回帰モデルに組入

擬似集団はどんな標的集団？

34

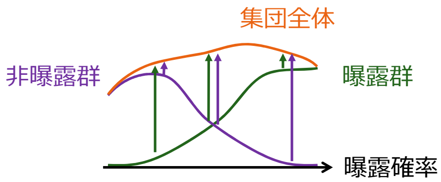
- 単純にマッチングした場合
- 「PSマッチングできた集団」



IPW解析での擬似集団①

35

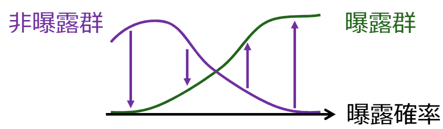
- 重みの作り方で、いろいろな集団に
  - Average treatment effect
    - 曝露群では曝露確率の逆数
    - 非曝露群では非曝露確率の逆数



IPW解析での擬似集団②

36

- 重みの作り方で、いろいろな集団に
  - Average treatment effect as treated (exposed)
    - 曝露群はそのまま
    - 非曝露群では曝露確率のオッズ



PS解析の流れ

37

- PSの推定
  - ロジスティック回帰、層別解析、決定木等
- PSの分布を確認、PS解析の方針
  - IPW(Proc CAUSALTRT)、マッチング(Proc PSMATCH)、層別、回帰
- 背景情報のバランス評価
  - 標準化差などで背景のバランスを評価
- 平均因果効果の推定

  ＋感度解析

因果推論に必要なもの

38

- 因果ネットワークに関する専門家の意見 と 検証不能な仮定
- 因果ダイアグラム causal diagram
  - 因果関係を仮定、図示化
  - 生じうるバイアスを整理
    - 交絡バイアス、選択バイアス、情報バイアス
  - 因果効果の分離
    - 直接効果・間接効果

有向非循環グラフ

39

- Directed Acyclic Graphs; DAGs
- Directed 有向
  - ノード node 間の矢線 arrow で順序性をいう
  - LがAの原因
- Acyclic 非循環
  - 自分自身の原因となることがない

$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A \arrow[r] & Y \end{tikzcd}$$

DAGで出てくる用語

40

- ノード、節点 node、点 vertex
  - 各変数をノードにおく
- 矢線 arrow、辺 edge
  - 一般に、辺は方向によらず使える言葉
- パス、経路、道 path
  - あるノードから異なるノードまでの行き方
- 親 parent
  - 祖先 ancestor：親の親、その親・・を含める
- 子 child
  - 子孫 descendant：子の子、その子・・を含める

因果DAG

41

- 以下のようなDAG
  - ノード間を直接結ぶ矢線がない場合、直接(因果)効果がない
    - あるかもしれない、なら矢線を示しておく
  - ある変数達に共通する原因は、観察できないとしても、同じグラフ上に示す
  - いかなる変数もその子孫に対し原因となる
- 因果DAGは背景にある反事実モデルを表現

周辺独立 marginally independent

42

- 因果DAGにおける2変数間の特徴
- 以下のいずれかを満たせば"(周辺)関連"
  - 一方がもう一方の原因
  - 共通の原因（親）をもつ
- 関連しない場合、(周辺)独立

$$\begin{tikzcd} L \arrow[r] \arrow[rr, bend left] & A & Y \end{tikzcd} \qquad \begin{tikzcd} A \arrow[rr, bend left] & Y \arrow[r] & C \end{tikzcd}$$

AとYは関連　　AとYは独立

条件付き独立 conditional independence

43

$$\begin{tikzcd} A \arrow[r] & \boxed{M} \arrow[r] & Y \end{tikzcd}$$

- AとYに周辺関連がある
  - Mは中間変数、媒介変数 mediator
- Mの水準を限定したら？
  - 条件付ける conditional on
  - □で囲う
- Mで条件付けることで、関連のあったパス$A \to M \to Y$をブロック
  - 条件付き独立にした

共通原因をBlocked

44

$$\begin{tikzcd} \boxed{L} \arrow[r] \arrow[rr, bend left] & A & Y \end{tikzcd}$$

- AとYに周辺関連がある
  - Lが共通原因
- Lを条件付け
- 関連のあったパス$A \leftarrow L \to Y$をブロック
  - 条件付き独立にした

合流点 collider をブロック

45

$$\begin{tikzcd} A \arrow[rr, bend left] & Y \arrow[r] & \boxed{C} \end{tikzcd}$$

- AとYは周辺独立
  - $A \rightarrow C \leftarrow Y$というパスは関連を生まない
  - Cが合流点 collider
- Cを条件付け
- $A \rightarrow C \leftarrow Y$という
  ブロックされていたパスをオープンに
  - 関連が生じる

合流点の子孫をブロック

46

$$\begin{tikzcd} A \arrow[rr, bend left] & Y \arrow[r] & C_1 \arrow[r] & \boxed{C_2} \end{tikzcd}$$

- 合流点のみならず、その子孫についても
  AとYは原因となっていた
- C_2で条件つけても、
  $A \rightarrow C_1 \leftarrow Y$をオープンに
  - 直接の合流点C_1で条件付けることと同様

blockedかopenか

47

- パスがblockedな状況は以下のいずれか
  - 非合流点で条件付け
    - 中間変数や共通原因で条件付け
  - 合流点とその子孫は条件付けない
- blockedでないパスがopen path

有向分離 d-separation

48

- 次の条件のいずれかを満たすとき、
  $\{A, Y\}$と排反な変数集合Sが
  $A - Y$間を有向分離する
  - $A - Y$間のすべてのパスにおける合流点で、
    その合流点と子孫がSに含まれないものがある
  - $A - Y$間のすべてのパスに非合流点で、
    Sに含まれるものがある
- Sで条件付ければ、
  $A - Y$間をつなぐパスをすべてblocked
  - open pathが含まれる場合をd-connected

練習④　有向分離するSは？

49

$$\begin{tikzcd} A \arrow[dr] & & L_2 \arrow[dl] \arrow[dr] & & Y \arrow[dl] \\ & L_1 & & L_3 & \end{tikzcd}$$

1) $\{\phi\}$　（空集合）
2) $\{L_1\}$
3) $\{L_1, L_2\}$
4) $\{L_1, \quad L_3\}$
5) $\{L_2\}$
6) $\{L_2, L_3\}$
7) $\{L_3\}$
8) $\{L_1, L_2, L_3\}$

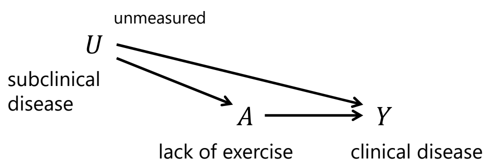
バックドア基準 back-door criterion

50

- AはYの非子孫
- 次の2条件を満たす頂点集合Sは
  $A - Y$についてバックドア基準を満たす
  - AからSの任意の要素へ有向道がない
  - Aから出る矢線をすべて除いたグラフにおいて、
    SがAとYを有向分離する
- S, A, Yが観察されていれば、
  AからYへの因果効果は識別可能

交絡の例①

51

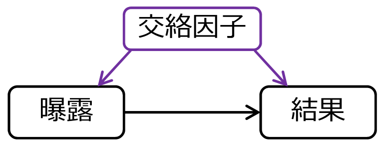


- 逆因果 reverse causation
- $A \leftarrow U \rightarrow Y$のバックドアパスが存在
  - Uをもし観察できれば、条件付けることで
    バックドアパスをブロック

交絡因子の経験的同定基準

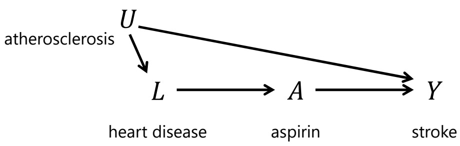
52

- 結果に影響を与える
- 曝露の有無によって分布が異なる
- 曝露から影響を受けない



交絡の例②

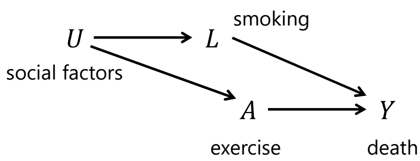
53



- 適応による交絡 confounding by indication
- UかLを条件付ければ$A \leftarrow L \leftarrow U \rightarrow Y$をブロック
  - LはUを介してYに影響するため、
    経験的同定基準でも交絡因子として定義される

交絡の例③

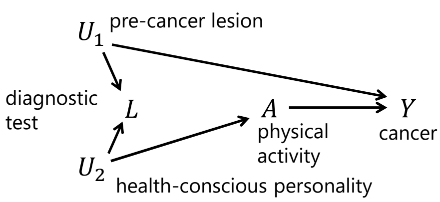
54



- UかLを条件付ければ$A \leftarrow U \rightarrow L \rightarrow Y$をブロック
  - LはUを介してAに影響するため、
    経験的同定基準でも交絡因子として定義される

交絡の例④

55



- $A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$は既にブロック
  - Lが合流点ゆえ
  - しかしLは経験的同定基準では交絡因子
- Lで条件付けるとバックドアパスが開く

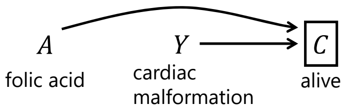
交絡が生じるかをDAGで

56

- 経験的同定基準の修正も考えられている
  - Greenland S, Pearl J, Robins JM. Epidemiology. 1999.
- バックドア基準で条件付けすべきか
  考えるほうがわかりやすい

選択バイアスの例①

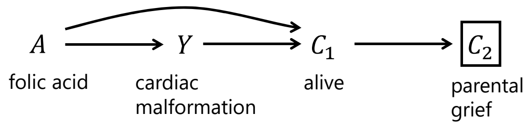
57



- 周産期の疫学研究でよくある例
  - 流産、死産例を無視することで生じる

選択バイアスの例②

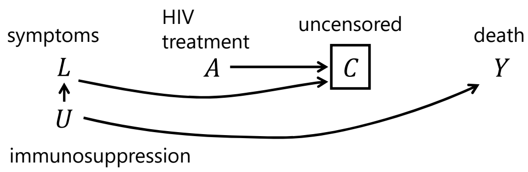
58



- 合流点の子孫で条件つけても同様

選択バイアスの例③

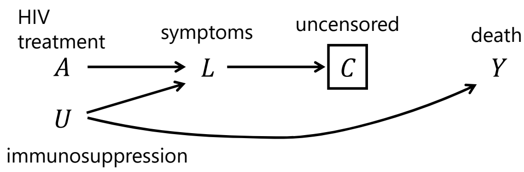
59



- Cを条件付けると
  $A \rightarrow C \leftarrow L \leftarrow U \rightarrow Y$のパスが開く

選択バイアスの例④

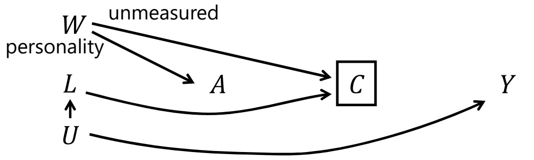
60



- 治療によって症状が変化し、
  観察継続されやすさが変化
  - 合流点の子孫で条件付けても同様

選択バイアスの例⑤

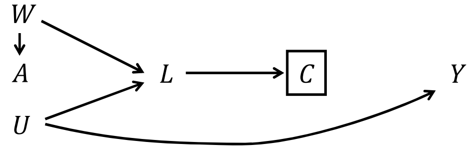
61



- ③のDAGにWを追加

選択バイアスの例⑥

62

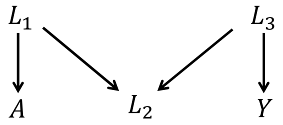


- ④のDAGにWを追加
- ⑤、⑥はMバイアスの一例

練習⑤　Mバイアス

63

- 有向分離する変数集合は？



1) $\{\phi\}$　（空集合）
2) $\{L_1\}$
3) $\{L_1, L_2\}$
4) $\{L_1, \quad L_3\}$
5) $\{L_2\}$
6) $\{L_2, L_3\}$
7) $\{L_3\}$
8) $\{L_1, L_2, L_3\}$

選択バイアスの種類

64

- Differential loss to follow-up
  - Missing data, Nonresponse bias
- Self-selection bias, Volunteer bias
  - Healthy worker bias
- Incidence-prevalence bias, Neyman bias
- Length time bias
- ...etc

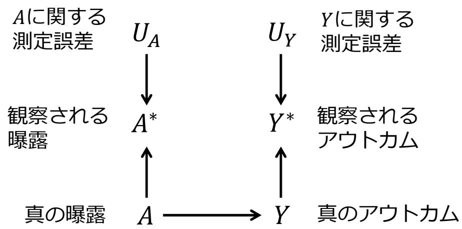
選択バイアスの対処

65

- デザインでの制御
- Biased samplingへの対処法を利用
  - IPW解析
    - 選択される確率の逆数で重みつけ
  - 標準化

情報バイアス

66



- $\frac{\Pr[Y=1|A=1]}{\Pr[Y=1|A=0]} = \frac{\Pr[Y^{a=1}=1]}{\Pr[Y^{a=0}=1]} \neq \frac{\Pr[Y^*=1|A^*=1]}{\Pr[Y^*=1|A^*=0]}$

測定誤差の特徴

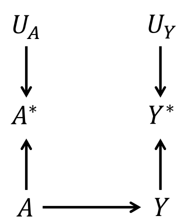
67

- 独立性 independence
- 非差異性 nondifferentiality

情報バイアスの例①

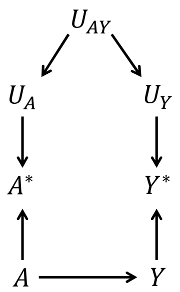
68

- independent
- nondifferential



情報バイアスの例②

69

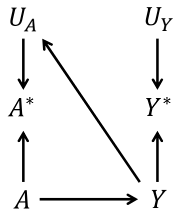


- dependent
- nondifferential

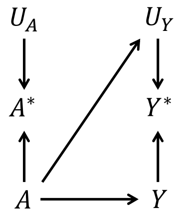
情報バイアスの例③④

70

③ recall bias



④ 非盲検の場合

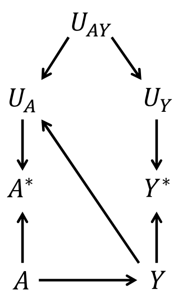


- independent differencial

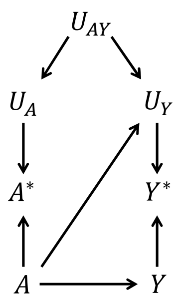
情報バイアスの例⑤⑥

71

⑤



⑥



- dependent differencial

independent, nondifferential

72

- 曝露が2値である場合、
  $A-Y$の関連に比べ、
  A^*-Y^*の関連は帰無に向かうバイアス
- 補正方法も検討されてきた
  - dependentやdifferentialな場合は難しい
  - むしろ、デザインでの制御がよいだろう

まとめ

73

- 反事実アウトカムと因果効果
  - 交換可能性と交絡
  - 標準化（層別解析）
  - 回帰モデルによる調整
  - 逆確率重み付け法
  - 傾向スコア
- 因果DAG
  - 有向分離 d-separation
  - 交絡バイアス、選択バイアス、情報バイアス

教科書など

74

- Hernán MA, Robins JM. Causal Inference. Boca Raton: Chapman & Hall/CRC, forthcoming.
  - HernanのWebより草稿を閲覧可能
- 黒木学. 構造的因果モデルの基礎. 共立出版. 2017.
- Rothman KJ. Greenland S. Lash TL. *Modern Epidemiology 3rd ed.* LWW. 2008.
- 田栗正隆. SASによる因果推論：CAUSALTRT プロシジャの紹介. 第36回SASユーザー総会. 2017/8/6.
  - Webより資料をDL可能