2019/10/7 北大・医理工統計学①

疫学研究と交絡

北海道大学　医学統計学
横田　勲

講義予定

2

1. 疫学研究と交絡
2. 因果DAG
3. 効果の修飾と層別解析
4. 傾向スコア；マッチングと重み付き解析
5. 回帰モデル入門
6. ロジスティック回帰とCox回帰
7. 予測モデルと診断法の評価
8. 臨床試験

成績評価

3

- 毎回のレポート
- 講義内での小テスト

今回の内容

4

- 生物統計学とその周辺的学問分野
- 疫学研究の例と交絡
- 反事実アウトカム

到達目標

- 生物統計学・疫学・臨床試験・機械学習の関係性を知る
- 疫学研究における交絡という現象を理解する
- 反事実アウトカムモデルを知る
- 標準化を行える

統計学

5

- もともと国勢を検討するための学問
- 帰納的な考え方
  - 演繹的な議論を展開する数学からすれば異端
  - 最近は「データサイエンス」
  - 不確実性を扱うため確率を用いて結果を示す
- 数字自体の意味を常に考える
  - 単位を意識する

データサイエンス

6

- データから何か知見を得るための学問
  - 統計学は大きく貢献
  - 計算機科学、情報工学、機械学習なども貢献
- 頻度流統計学の思想
  - データ生成過程をもとに仮説をおき、得られたデータが仮説に従うかを判断する
- データサイエンスの思想
  - データから予測された内容が当たり、利益をもたらせばよしとする

手元にあるデータをどう活用するか 7

- 記述統計
  - どのようにデータが得られたかを明らかに
  - データのタイプに応じた要約
    - 連続量、カテゴリカル、生存時間
- 推測統計
  - 想定する源泉集団において、曝露や治療とアウトカムの関連を検討
  - 因果関係を明らかにしたい、予測を行いたい
  - 統計的検定、推定を利用

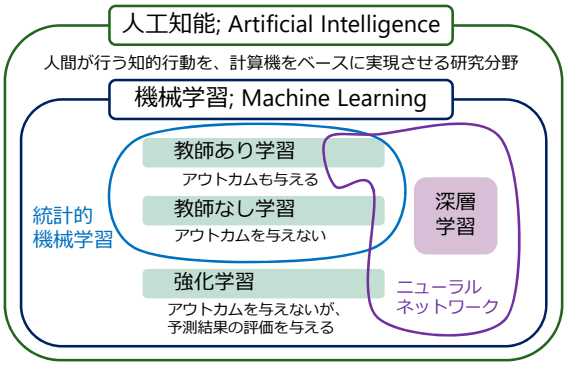
生物統計学 Biostatistics 8

- 「生き物」を扱う分野に応用する統計学
  - 特にヒトは異質性（個体差）が大きく、なかなか計画通りにデータを測定できないし単純な確率分布からデータは生成されない
- 臨床試験や疫学研究の計画・解析の方法論・理論的基礎
  - データをどう取るかが最も重要

ビッグデータ 9

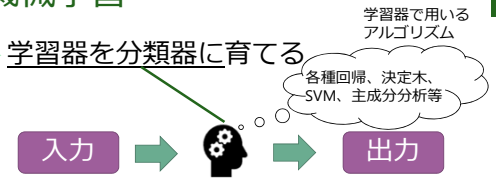
- 医療データベース
  - レセプトデータ
  - 電子カルテデータ
  - ・・・
- マーケティングデータ、気象データ、、
- 色々なデータが利用可能になってきた
  - ICT技術の発達により桁の違うデータ数

人工知能と機械学習 10



機械学習 11

- 学習器を分類器に育てる



  - 入力に様々な訓練データを与える
    - 出力に正解（教師データ）を与えて、入力から出力になるべく近い結果を導くよう訓練
    - 出力には何も与えず、入力の特徴を抽出して、分類結果を出力

12

https://products.sint.co.jp/aisia/blog/vol1-13

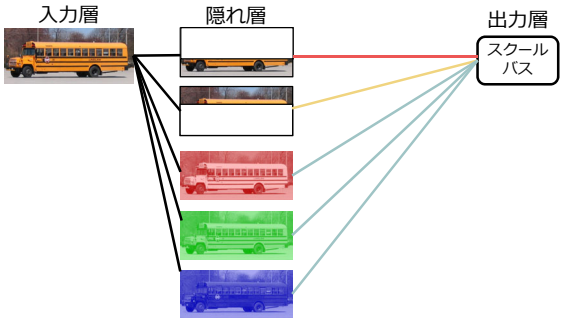
教師あり v.s. 教師なし

13

- 教師あり
  - 目的に合った精度の高い分類
- 教師なし
  - 教師つき学習用データを用意する必要がない
    - 一昔前までは教師あり、よりも膨大なデータ数が必要
  - 予想外の結果が得られやすい
- 半教師つき学習
  - 最初だけ教師ありで、
    途中から膨大な教師なしデータで学習

ニューラルネットワーク

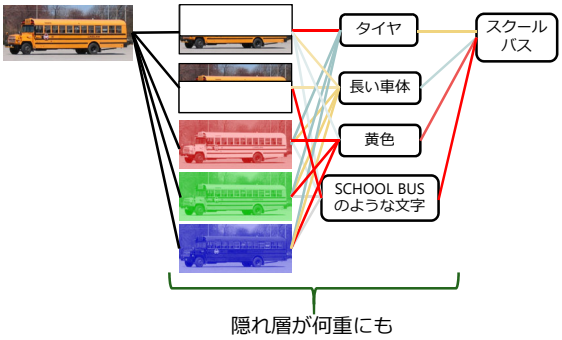
14



深層学習

15

本当のところ、
どのような隠れ層になっているかは
全くわからない



人間 v.s. AI

16

- 1996-7 チェス
- 1997 オセロ
  - 1秒間で約$10^{6\sim8}$通りの手を読めていた
- 2013-5 将棋
  - ここまでは将来の手の全数検索
- 2017 囲碁
  - 囲碁の手数は10^{172}通り https://tromp.github.io/go/legal.html
  - 深層学習を用いたAlphaGo

会話用AI「Tay」

17

- 人間との会話に特化させたAI
  - 19歳の米国人女性、という設定
  - Twitterを介して会話を学習
- 2016年 Microsoft社が開発
  - 3/23 公開
  - 3/25 差別的発言を繰り返すため一時停止
  - 3/30 復旧→違法行為をツイート→即終了
- 入力（学習内容）が酷かったことが一因

AIの限界

18

- データから学習するしかない
  - データに潜むあらゆる不正確性が
    そのまま結果に反映
  - 予測や分析などの追加機能は個別に追加
    https://www.sas.com/ja_jp/insights/analytics/what-is-artificial-intelligence.html
- 学習させるデータの質が重要
  - アルゴリズムの選択より重要といわれる

疫学 epidemiology の視点が大切

John Snowとコレラ

19

John Snow. *On the Mode of Communication of Cholera*. 1855

- 1854年ロンドンのBroad St.でコレラ発生
  - 半径250ydの所で10日間で500人以上が死亡

住民の1割以上

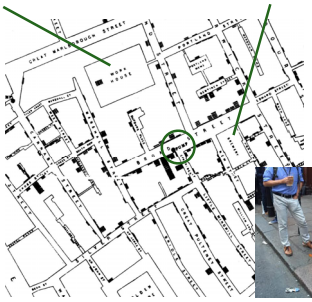
ポンプの位置と死亡者を地図上にプロット

あるポンプに注目

20

救貧院では535人中5人(1%以下)のみ死亡
・ポンプで水は汲まず、構内井戸を使用していた

醸造所社員70人中死亡者ゼロ
・水は飲まずにビールを飲んでいた

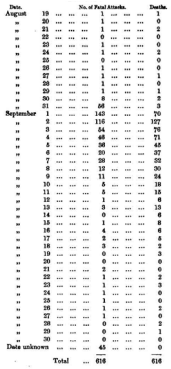


本物ポンプはロンドン大学熱帯衛生医学大学院に展示

当時の縁石が1枚残されている

各日の発生数と死亡数

21



9/8にポンプの柄を除去

Lambeth社とSouthwark&Vauxhall社

22

- どちらも同じような水の汚さ
  - テムズ川の下流から水をひいていた
  - 水洗便所普及により下水は川に垂れ流し
- 顧客の大半は庶民
- 1852年、Lambeth社 取水地を上流に

コレラ死亡率の比較

23

	契約戸数	コレラ死亡者数	1万戸あたり死亡
S & V社	40,046	1,263	315
Lambeth社	26,107	98	37
その他	256,423	1,422	59

- Southwark & Vauxhall社のほうが8.5倍の死亡リスク

水を介して感染する

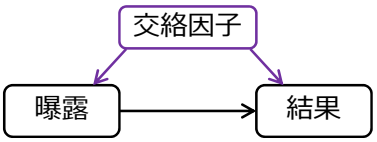
24

- 悪臭（空気感染）説
  - 同じ部屋にいただけで感染しないよ
- 標高説
  - 標高が高い地域ではコレラが少ない
  - 水供給の違いで説明
- 水道会社の選択とコレラのなりやすさは独立（交絡がない）
  - 一方の会社が取水地を変えたら死亡者減
  - ポンプ栓を抜いたらコレラ発生者減

交絡 confounding

25

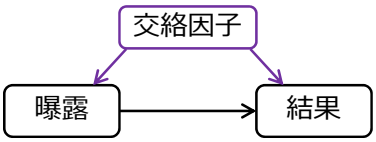
- 曝露による効果を調べる際に、曝露が結果に影響する他の因子と関連することで、歪められた効果が観察されてしまう現象
  - 交絡因子：交絡をひきおこす因子



交絡因子の定性的必要条件

26

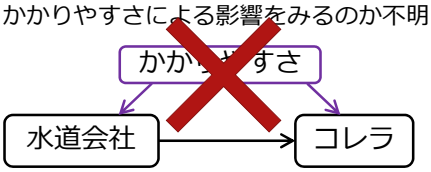
- 結果に影響を与える
- 曝露の有無によって分布が異なる
- 曝露から影響を受けない



コレラの例

27

- もしコレラにかかりやすい人が、特定の水道会社を選んでいたら・・？
  - L社はコレラにかかりにくい人ほど選ぶ
  - S&V社はコレラにかかりやすい人ほど選ぶ
  - 水道会社の比較では、水質の影響をみるのか、かかりやすさによる影響をみるのか不明



手術成績の比較（仮想例）①

28

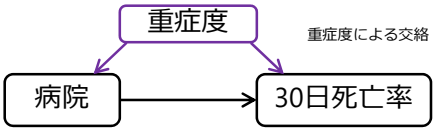
- 30日死亡率を比べてみよう
  - 専門医がたくさんいる病院A：2%
  - 一般的な病院B：1%

この結果だけから、あなたは病院Bを選びますか？

手術成績の比較（仮想例）②

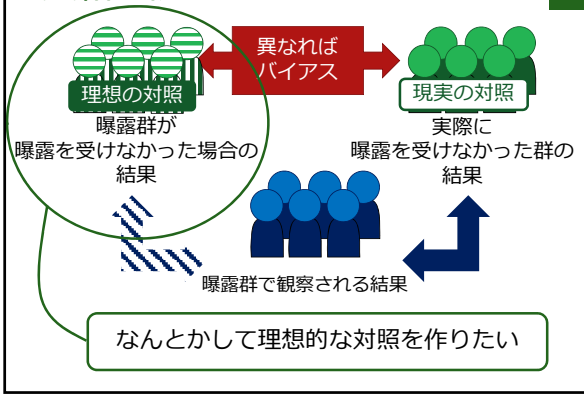
29

- 30日死亡率を比べてみよう
  - 専門医がたくさんいる病院A：2%
  - 一般的な病院B：1%
- 病院Aは近隣の医療機関から紹介をたくさん受けている
  - より重症な人ほど病院Aにかかりやすい



交絡を防ぐには

30



なんとかして理想的な対照を作りたい

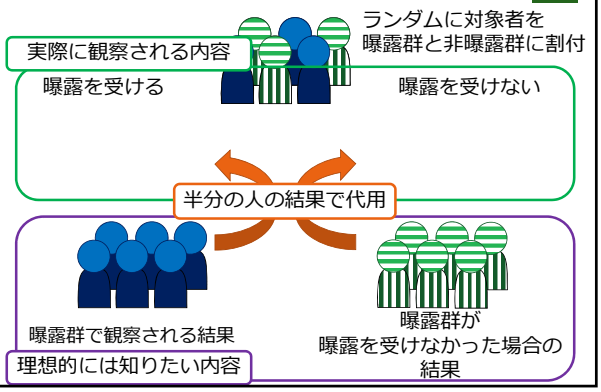
個人では基本的にムリ

31

- 曝露を受けた場合の結果を知った後に曝露を受けさせないことはできない
- 集団で議論することはできないだろうか

ランダム化 randomisation

32



脚気論争

松田誠. 慈恵医大誌. 2002, 2003.

33

- ビタミンB1欠乏症
  - 白米ばかりの食生活
  - 大正時代は結核と並ぶ二大国民病
- 高木兼寛 (1849-1920)
  - William Willisより
    イギリス医学を学ぶ
  - 海軍軍医になり、
    イギリス留学中(1875-80)
    疫学の方法論を学ぶ
  - 帰国後、脚気調査に

栄養欠陥説

34

- 龍驤艦　明治15(1882)～16年
  - 378名中169名が脚気に罹患、25名死亡
    - 外国で停泊中は脚気患者が減る
    - 寄港したハワイで食料を全部入れ替えて以降、誰も罹患しなかった
  - タンパク質を多くすれば予防できるのでは
- 筑波艦　明治17年
  - 食事を洋食に変え、龍驤艦と同一航路
  - 333名中14名が罹患、死亡者ゼロ
    - 14名のうち12名は洋食を拒んだ

海軍は麦食に

35

- 洋食はウケが悪かった

年次	食事	兵員数	脚気患者数	死亡者数
明治11(1878)年	米食	4,528	1,485	32
明治12年	米食	5,030	1,978	57
明治13年	米食	4,956	1,725	27
明治14年	米食	4,641	1,163	30
明治15年	米食	4,769	1,929	51
明治16年	米食	5,346	1,236	49
明治17年	洋食	5,638	718	8
明治18年	麦食	6,918	41	0
明治19年	麦食	8,475	3	0
明治20年	麦食	9,016	0	0
明治21年	麦食	9,184	0	0

栄養欠陥説への批判

36

- 緒方正規「脚気病菌発見」(1885)
  - 北里柴三郎により実験不備を指摘され消える
- 大沢謙二「食物消化の試験」(1887)
  - 栄養吸収は米のタンパクのほうがよい
- 森林太郎「統計二就テノ分疏」(1889)
  - 患者の減少時期と麦食への切り替え時期が偶然一致しただけ
  - スタチスチックは科学でなく方法であり、原因を探り法則を知り得るものではない

ランダム化比較試験の示唆 37

- 一つの兵団を二分して、
  一方に麦食を、もう一方に米食を与え、
  両者を同一の地に住まわせ、
  他の生活条件も同じにすべき
- もし米食者のみが脚気に罹り、
  麦食者は罹らなかったら、米食が原因だ

森林太郎「統計二就テノ分疏」(1889)

日清戦争から日露戦争 38

- 日清戦争(1894-95)
  - 陸軍：40,000+の罹患、4,000+の死亡
    - 戦死者は約300
  - 海軍：罹患ゼロ
- 義和団の乱(1900)
  - 第五師団20,000+中、2,351の罹患
- 日露戦争(1904-05)
  - 陸軍：250,000+の罹患、約28,000の死亡
    - 戦死者は約47,000
  - 海軍：若干の罹患、死亡ゼロ

疫学的観点から 39

- 脚気論争
  - 高木の栄養欠陥説自体は間違っていた
  - 食事に問題があることは正しかった
    - 疫学研究によって特定できた
    - 航海実験から約30年後の1911年、
      鈴木梅太郎によるビタミンB1発見
- John Snowとコレラ
  - 水に問題があることを疫学研究で特定
    - メカニズム解明は、約30年後の1884年、
      Robert Kochによるコレラ菌再発見
      （発見自体は1854年Filippo Paciniによる）

反事実アウトカム 40

- 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さんが米食の船に乗って（$a_H = 0$）、
    脚気になった（$Y_H = 1$）
- 反事実データ counterfactual data
  - Hさんが洋食も出る船に乗ったら（ $a_H = 1$ ）、
    脚気になったか？（Y_H =?）

潜在アウトカム potential outcome 41

- $Y^{a=0}$
  - 曝露$a = 0$を受けた場合のアウトカム
- $Y^{a=1}$
  - 曝露$a = 1$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

	$Y^{a=0}$	$Y^{a=1}$
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 42

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

個人での因果効果 43

	$Y^{a=0}$	$Y^{a=1}$	Causal effect $Y^{a=1} - Y^{a=0}$
Doomed	1	1	$1-1=0$
Helped	1	0	$0-1=-1$
Hurt	0	1	$1-0=1$
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 44

- $E[Y^{a=1}] - E[Y^{a=0}]$
  - 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 45

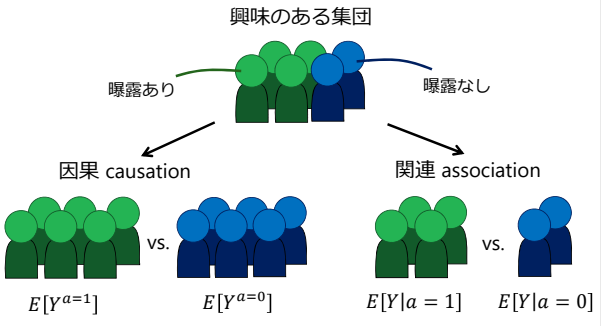
- 因果リスク差
  - $\Pr[Y^{a=1}=1] - \Pr[Y^{a=0}=1]$
- 因果リスク比
  - $\dfrac{\Pr[Y^{a=1}=1]}{\Pr[Y^{a=0}=1]}$
- 因果オッズ比
  - $\dfrac{\Pr[Y^{a=1}=1]/\Pr[Y^{a=1}=0]}{\Pr[Y^{a=0}=1]/\Pr[Y^{a=0}=0]}$

練習① 以下のACEは？ 46

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 47



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

関連効果の指標 48

- 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 49

- 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 50

- 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 51

exchangeability

- 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 52

- 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 53

- 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 54

- 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？

55

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

おわりに

56

- 生物統計学・疫学・臨床試験・機械学習の関係性を知る
- 交絡という現象を理解する
- 反事実アウトカムモデルを知る
- 標準化を行える