大阪大学基礎工学部公開講座
日程：平成18年8月28日～31日
於：　基礎工学部国際棟

リテラシーとしてのデータ科学
--- 原因と結果の妙 ---

システム創成専攻 数理科学領域
狩野 裕

リテラシーとは

- リテラシー
  - 広辞苑（第五版）によると
    - 読み書きの能力。識字。転じて、ある分野に関する知識・能力。「コンピューター-リテラシー」
- Literacy
  - Cambridge Learner's Dictionary
    - the ability to read and write
- 本講義の目的
  - データの見方についてのリテラシー
  - 特に関係・連関性について

本講義の内容

1. 相手がいる
   - 関西には「たこ焼き器」が多い
   - スペースシャトル爆発事故
2. 相手がいても
   - 婚姻率と死亡率
   - 子のつく名前の女の子は頭がいい
3. 統計的因果推論
   - 反事実モデル
4. 結語

1. 相手がいる

関西には「たこ焼き器」が多い

たこ焼き器

- 関西にはどの家にも「たこ焼き器」がある
- 関西には「たこ焼き器」が多い

$$\frac{34}{39} = 0.87 = 87\%$$

- 関西には関西以外と比べて「たこ焼き器」が多い

$$\frac{14}{36} = 0.39 = 39\%$$

育成地	たこ焼き器		
	あり	なし	合計
関西	34	5	39
非関西	14	22	36
合計	48	27	75

2×2の分割表（クロス集計表）

- 0.87と0.39は大きく異なる
- たこ焼き器の有無は関西かどうかと関係
  - 関西にはたこ焼き器が多いと言ってよい
- 二つの属性（変数）が無関係（独立）

$$\Longleftrightarrow \frac{a}{a+b} \approx \frac{c}{c+d}$$

育成地	たこ焼き器		
	あり	なし	合計
関西	a	b	a+b
非関西	c	d	c+d
合計	a+c	b+d	a+b+c+d

分割表の見方

- 「多い」「少ない」という表現は何かと比較して言うことが普通である
- 関西で87%のお宅にたこ焼き器があったとしても，それだけでは，「関西に多い」という主張はできない
- 非関西と比較して初めて「より多い」ということが主張できる

スペースシャトル爆発事故

スペースシャトル

- 8/30「アトランティス」打上げ予定
  - 9月初旬まで延期．ハリケーン接近のため
- 7/18「ディスカバリー」が帰還
- 2003/2「コロンビア」空中分解事故
- 1986/1「チャレンジャー」爆発事故

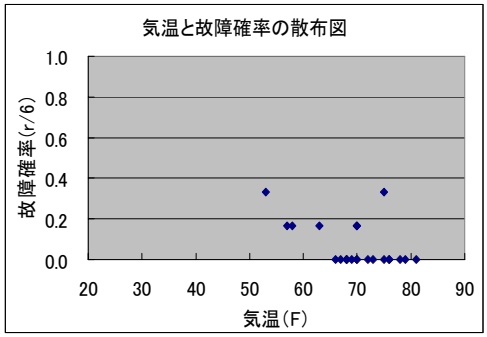
1986年スペースシャトル チャレンジャー号爆発事故

- 事故調査班は原因を「O-リング」という部品の故障だと断定
- 調査班は事故につながる重要な要因として（外）気温を取り上げている
- 過去のデータに基づき，打上時の気温（31°F）から故障確率を予測すると？

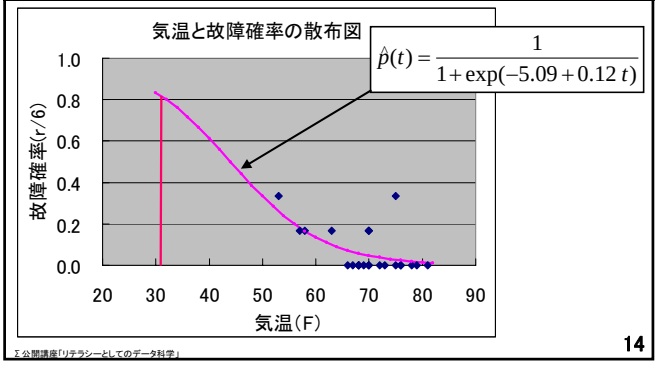
過去23回のスペースシャトル打ち上げ時の気温と「O-リング」故障数（全6個中）

故障数	気温	故障数	気温	故障数	気温	故障数	気温
2	53	1	58	0	73	0	76
0	70	1	70	0	67	0	70
0	78	0	81	0	75	0	76
1	57	1	63	0	68	0	68
1	70	0	72	2	75	0	68
0	79	0	66	0	69		

気温と故障確率の散布図



推定された故障確率
by ロジスティック回帰分析



回帰式の利用
チャレンジャー号が爆発したとき(t=31)の故障確率は？

- 推定された回帰式

$$\hat{p}(t) = \frac{1}{1+\exp(-5.09+0.12\,t)}$$

- 気温が31°Fでの故障確率の点推定値

$$\hat{p}(31) = 0.82$$

- 6つの「O-リング」のうち少なくとも1つが故障する確率

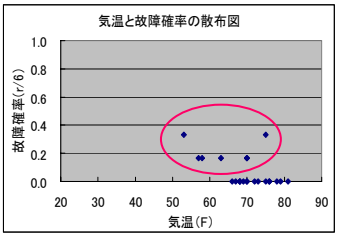
$$1-(1-\hat{p}(31))^6 = 0.99996$$

ここまでのまとめと疑問

- 過去のデータから，t=31°Fで打ち上げたとき，O-リングが故障する確率は非常に高いことが予想できる
- NASAは気温とO-リングの故障との関係を検討していた
- では，なぜ，打ち上げに踏み切ったのか？

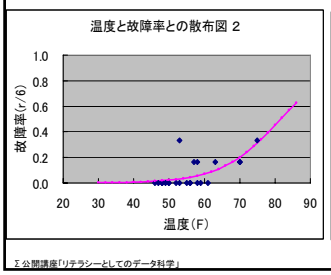
NASAの技術者は．．．

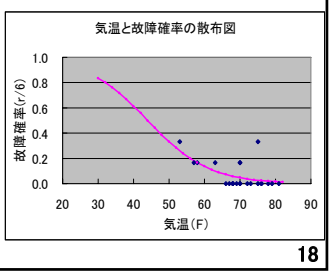
- 「故障した打ち上げデータ」だけを吟味していた
- 気温と故障の関係は適切に同定できない！



たとえば．．．

- もし正常打ち上げのデータが．．．．





まとめ

- 関係性を検討したいときは「相手」が必要
  - 関西にはたこ焼き器が多い
    - 非関西では？
  - O-リング故障の原因は外気温
    - 故障しなかったデータでは？

2. 相手がいても

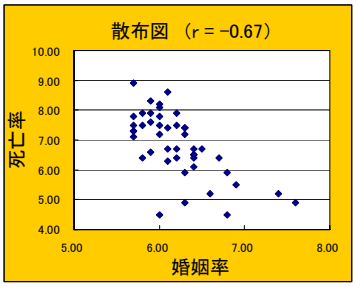
婚姻率と死亡率
偏相関係数

婚姻率と死亡率

偏相関係数

婚姻率と死亡率

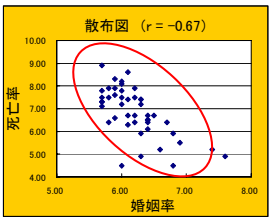
	死亡率	婚姻率
石　川	6.6	5.9
福　井	7.2	6.3
山　梨	7.8	6.0
長　野	7.3	5.7
岐　阜	6.4	5.8
静　岡	6.1	6.4
愛　知	5.2	6.6
三　重	7.2	6.0
滋　賀	6.5	6.4
京　都	6.4	6.2
大　阪	5.5	6.9
兵　庫	6.4	6.4
奈　良	6.3	6.1
和歌山	8.1	6.0

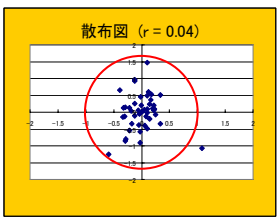


都道府県別のプロット
出典：人文・社会科学の統計学（東大出版）

相関係数

- 相関の正負と強さを記述
- −1 ≦ r ≦ 1

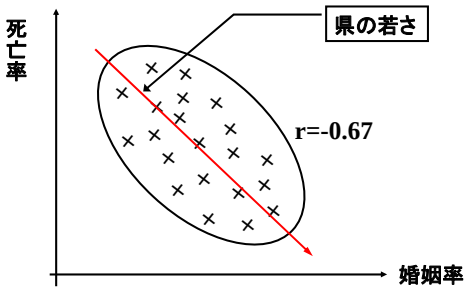




婚姻率と死亡率

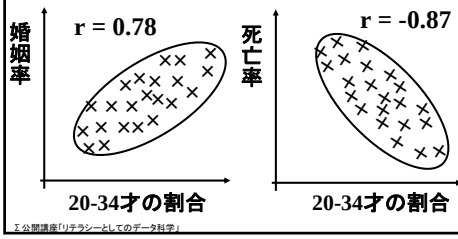
- 婚姻率と死亡率との間に明らかな負の相関
- 婚姻率を上げると死亡率が下がるのか？
  - という訳でもなかろう
- 定義（人口千人当たり）
  - 婚姻率：ある年度に提出された婚姻届の数
  - 死亡率：ある年度に提出された死亡届の数
- →
  何か別の要因が絡んでいる

擬相関
(spurious correlation)



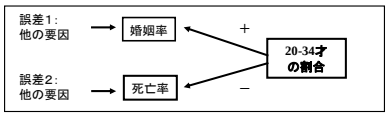
県の若さとの関係

・「県の若さ」を表す指標として，
「20-34才の割合」を考える



	死亡率	婚姻率	20-34歳の割合
石川	6.6	5.9	0.23
福井	7.2	6.3	0.21
山梨	7.8	6.0	0.21
長野	7.3	5.7	0.20
岐阜	6.4	5.8	0.22
静岡	6.1	6.4	0.23
愛知	5.2	6.6	0.25
三重	7.2	6.0	0.21
滋賀	6.5	6.4	0.23
京都	6.4	6.2	0.24
大阪	5.5	6.9	0.25
兵庫	6.4	6.4	0.23
奈良	6.3	6.1	0.23
和歌山	8.1	6.0	0.20

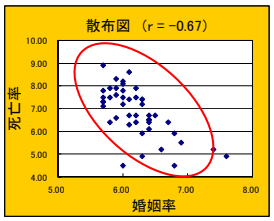
仮説モデルと検証

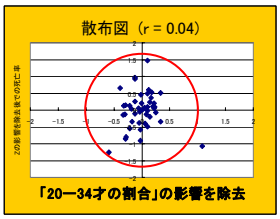


・「20-34才の割合」が婚姻率と死亡率を媒介するモデル
・「20-34才の割合」の影響を取り除くと婚姻率と死亡率との間の関係は消滅するはず

「20-34才の割合」の影響を取り除くと．．．

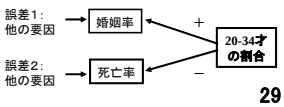
・偏相関係数を使う
・婚姻率と死亡率との間の負の相関は消滅！





まとめ

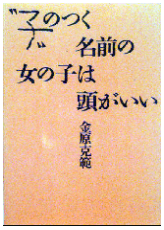
・婚姻率と死亡率との間の負の相関は直接的な関係ではない
・「県の若さ(20-34才の割合)」が婚姻率と死亡率を結ぶ
  − 交絡変数(第三変数，共変量)という
  − 交絡変数によって生じる関係を擬相関という
・婚姻率を上げても死亡率が下がる保証はない
・→
  若人の割合を上昇させれば婚姻率を上げ死亡率を下げることができる



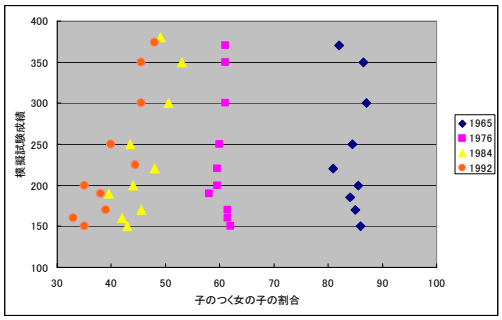
子のつく名前の女の子は頭がいい

名前と人生

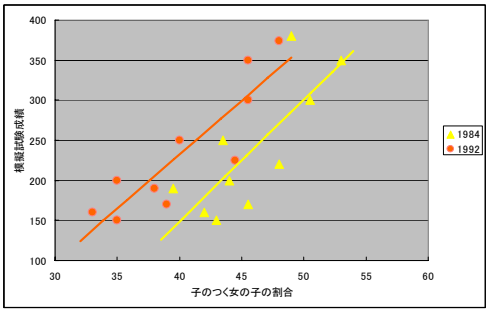
- 子のつく名前をつけると頭がよくなる？
- 命名と子供の特質に関係がある？
- 命名という行為によって子供に変化があるのは不自然



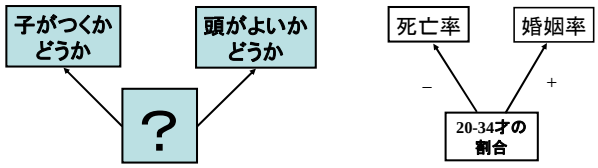
データは語る



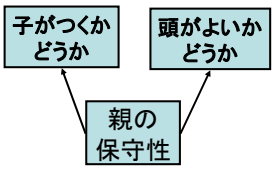
データは語る

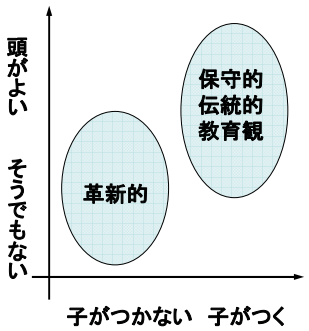


交絡変数の存在をうたがう



親の保守性

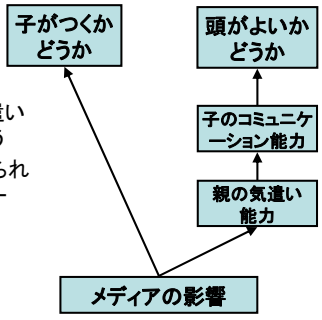




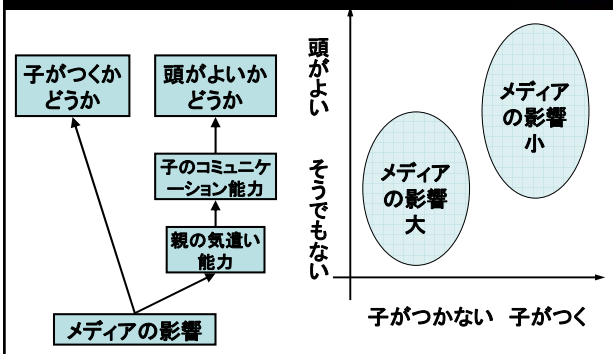
- 両親が保守的
  - 保守的な名前
  - 勉学中心の子育て

親へのメディアの影響

- 著者の主張
  - メディアは、人の気遣い能力を破壊してしまう
  - そのような親に育てられた子供はコミュニケーションに絶望する



親へのメディアの影響

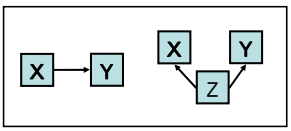


まとめ

- 「子のつく名前の女の子は頭がいい」というデータがある
- 命名によって頭がよくなるわけではない
- 擬相関である
  - 交絡変数として「親の保守性」「親へのメディアの影響」等が考えられる
  - この研究の意義は交絡変数（の候補）を見出したことにある
- 交絡変数を観測し偏相関係数を吟味したい

まとめ

- 二つの事柄間の関係
  - 因果＝「原因＋結果」
  - 交絡変数による擬相関



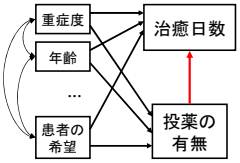
- 直感に反する関連には擬相関をうたがう
- 擬相関という用語にはマイナスのイメージ
  - 因果ではないという警鐘
  - 因果構造の一つであって，正確に理解し積極的に活用したい

最新の統計的因果推論

Rubinの因果

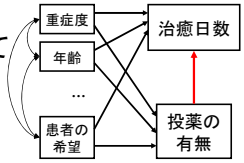
セットアップ

- 投薬の効果を調べる
  - 投薬群と非投薬群をつくる
  - 群間で治癒日数を比較
- 群のつくり方
  - 無作為割当て
    - 実験研究
  - 患者や医者の判断に委ねる
    - 観察研究



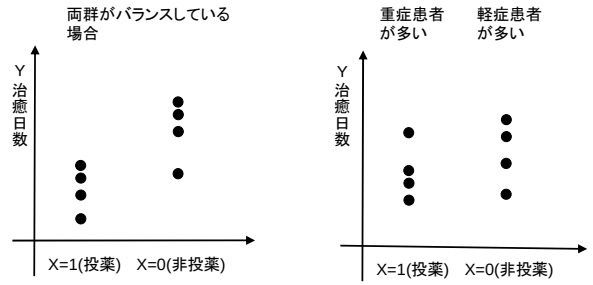
セットアップ

- 調査（or 実験）研究において
  - X：二値の原因変数
  - Y：結果変数
  - $\mathbf{z}=[Z_1,Z_2,\cdots,Z_m]'$：交絡変数



Y	Y_1	Y_2	$\cdots$	$\cdots$	Y_{n-1}	Y_n
X	0	$\cdots$	0	1	$\cdots$	1
$\mathbf{z}$	$\mathbf{z}_1$	$\mathbf{z}_2$	$\cdots$	$\cdots$	$\mathbf{z}_{n-1}$	$\mathbf{z}_n$

交絡変数の影響
投薬の効果があったとしても. . .



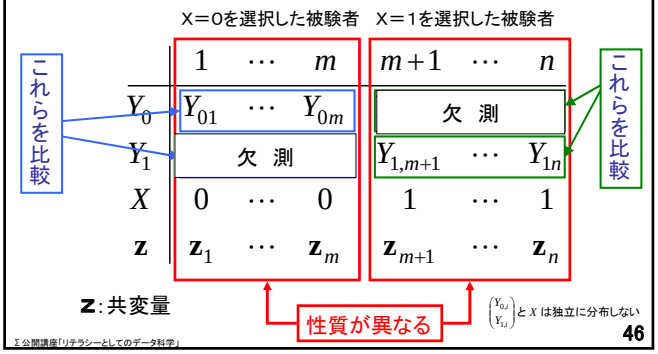
傾向スコア

- 傾向スコア(propensity score)
  - by Rosenbaum-Rubin (Biometrika, 1983)
  - $e(\mathbf{z})=P[X=1|\mathbf{z}]$
    - X=1を割付ける条件付確率
- 重要な性質
  - $X \perp\!\!\!\perp \mathbf{z} \mid e(\mathbf{z})$
  - $e(\mathbf{z})$は1次元

処方箋:
反事実モデルの導入

- 「投薬＋重症患者」と「非投薬＋軽症患者」の比較には意味がない
- ではどのように考えればよいか
  - 同じ患者について「投薬と非投薬」を比較すればよい
  - 必ず一方は欠測
  - 反事実モデル(仮定法過去)
    - もし「投薬していたとすれば」
      もし「投薬していなかったら」

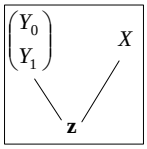
データの構造と欠測



重要な仮定:
強い意味で無視可能

- Strongly Ignorable
  - zを与えた下でXと(Y_0,Y_1)は条件付独立

$$\begin{pmatrix} Y_0 \\ Y_1 \end{pmatrix} \perp\!\!\!\perp X \mid \mathbf{z}$$



  - X は $\mathbf{z}$ からのみ直接的な影響を受ける
  - $\mathbf{z}$ を与えた下でバランスのとれた割付け
  - すべての交絡変数が $\mathbf{z}$ の中に含まれている
- この下で次式が成立

$$E[Y_i X|\mathbf{z}] = E[Y_i|\mathbf{z}] \times E[X|\mathbf{z}] = E[Y_i|\mathbf{z}]e(\mathbf{z})$$

欠測の母数を推定する

$$E[Y_1 X|\mathbf{z}] = E[Y_1|\mathbf{z}]e(\mathbf{z})$$

$$\therefore \quad E\left[\frac{Y_1 X}{e(\mathbf{z})}\middle|\mathbf{z}\right] = E[Y_1|\mathbf{z}], \qquad \therefore \quad E\left[\frac{Y_1 X}{e(\mathbf{z})}\right] = E[Y_1]$$

$$\therefore \quad E[Y_1|X=0] = E\left[\frac{(1-e(\mathbf{z}))Y_1 X}{e(\mathbf{z})P(X=0)}\right]$$

欠測の母数を推定する

$$E[Y_1|X=0] = E\left[\frac{(1-e(\boldsymbol{z}))Y_1X}{e(\boldsymbol{z})P(X=0)}\right]$$

$$\Longrightarrow$$

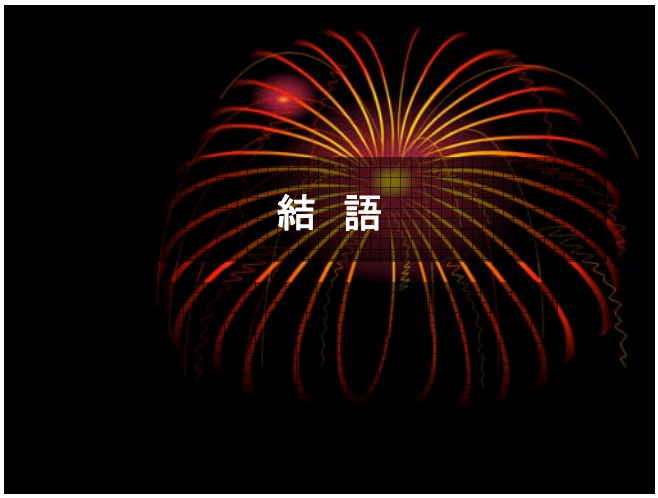
$$\widehat{E[Y_1|X=0]} = \frac{1}{n}\sum_{i=m+1}^{n}\frac{(1-e(\boldsymbol{z}_i))Y_{1i}}{e(\boldsymbol{z}_i)P(X=0)}$$

	X=0を選択した被験者			X=1を選択した被験者		
	1	⋯	m	$m+1$	⋯	n
Y_0	$E(Y_0 \mid X=0)$			$E(Y_0 \mid X=1)$		
Y_1	$E(Y_1 \mid X=0)$			$E(Y_1 \mid X=1)$		
X	0			1		
$\mathbf{z}$	$\mathbf{z}_0$			$\mathbf{z}_1$		

$$\approx \frac{\sum_{i=m+1}^{n}\frac{(1-e(\boldsymbol{z}_i))Y_{1i}}{e(\boldsymbol{z}_i)}}{\sum_{i=m+1}^{n}\frac{1-e(\boldsymbol{z}_i)}{e(\boldsymbol{z}_i)}}$$

まとめ

- 反事実モデル
  - 同一の被験者に二つの条件を割付けたとして比較する
    - 交絡変数の影響を殺すためのモデル
  - 一つの条件の下でのデータは欠測
    - 強い意味の無視可能の仮定の下で欠測値を推定
  - 幅広い応用があり，現在も発展中



結　語

- リテラシーとしてのデータ科学
  - データの適切な見方と利用方法を身に付ける
  - データに基づく理解しやすい議論を行う
  - 統計の誤用と悪用を避ける
- いくつかのポイント
  - 比較の重要性．．．相手が必要！
  - 擬相関．．．共通要因（交絡要因）による関連
- データの素性
  - どこで，誰を対象に，どのようにして得たデータか
  - 本講義ではふれることができず