リテラシーとしてのデータ科学

基礎工学研究科 システム創成専攻
数理科学領域
教　授　　**狩野　裕**

１．はじめに

私たちは「こうすればこうなる」「こういう人はこういうことをする」といったことをしばしば口にします．いくつか例を挙げましょう．

(1) 関西にはたこ焼き器が多い
(2) タバコを吸うと肺がんになり易い
(3) 婚姻率が高い都道府県では長生きできる
(4) 子のつく名前の女の子は頭がいい
(5) コウノトリが数多く飛来した年は出生数が多い
(6) 最新のケイタイをもつ人は男女別姓に賛成する
(7) 車で動物を轢いたなら，近々人身事故を起こす

私たちは，このようなことを，原因と結果，予測，相関・連関などと言います．研究活動や会社の業務の多くの場面で，因果関係や予測といったことが解決すべき問題点ではないでしょうか．

私たちがこのような問題・課題に出会ったとき，最初にすることは思考実験です．すなわち，その事実を正しいと考えるべきか疑うべきかを頭の中で色々な角度から検討するのです．科学的な研究やリスクの伴う判断に迫られた場合は，思考実験に止まらず，微分方程式などを用いて数理モデルを立てて分析したり，実験や調査を行い客観的なデータを採取し統計モデルを用いたりして，科学的証拠を得ようとします．本講では後者の場合を扱い，上述(1)-(7)のようなステートメントへの考え方や，データの見方についてお話しします．

以下の節ではいくつかの実例を紹介します．考え方のヒントだけを述べますので，講義までに皆さん自身で「思考実験」をしてみてください．

２．相手が必要

2.1 関西にはたこ焼き器が多い？

近年，たこ焼きと関西弁は全国区になったと言われています．では，たこ焼きを「焼く」という文化はどうでしょうか．つまり，関西ではたこ焼き器を持っている家庭が多いのでしょうか．表１のデータをみて考えてみましょう．

表１

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

2.2 スペースシャトル

1986 年,米国 NASA が打ち上げた有人飛行衛星チャレンジャー号が，茶の間で多くの米国人が見守る中，大爆発を起こして打ち上げに失敗しました．あのときの米国人が受けたショックは大変大きなものであったことは容易に想像がつきます．この事故の原因はシャトルにあるオー・リングという部品の故障でした． NASA の技術者は，当然ながら，打ち上げ前にオー・リングの問題を検討していました．図１はチャレンジャー以前の 23 回のシャトル打ち上げにおいて，オー・リングが故障した打ち上げ(n=6)における，オー・リングと外気温の関係をプロットした散布図で，NASA はこのデータを吟味していました．実は，ここには大きな落とし穴があったのです．

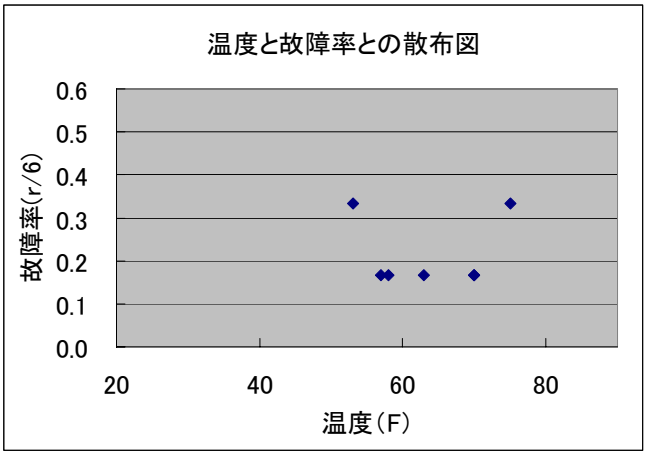


図１

３．相手がいても

この節ではもう少し複雑な関係を見る方法を紹介しましょう．

3.1 タバコと肺がん

統計的な方法で社会を動かした有名な例の一つにタバコと肺がんの関係があります．最近は日本でも健康増進法が施行される等，タバコと肺がんの関係を疑う人はいないと思いますが,1950 年ごろは,関係を否定するグループと肯定するグループが対峙し，タバコが肺がんの原因になるかどうかは大問題でした．統計学の基礎を築いた学者として有名な R. A. Fisher は否定派で，関係を支持する多くの相関研究を糾弾したこと

でも知られています．彼は愛煙家であり彼の判断にはバイアスがありましたが，データの見方という点では大いに参考にすべきところがあります．また，米国のタバコ会社も，当然ながら，否定派でした．彼らの主張は，図２のように，喫煙量 X と肺がん発症 Y の両者に関係する第三変数が存在して，それが X と Y を結んでいるという主張でした．すなわち，X と Y には関係があるように見えるのですが，実際は直接的因果関係はないというものでした．タバコ会社は第三変数として「ストレス」を主張しました．ストレスが高いと喫煙量が増えると同時に癌の発症確率も上昇する，その結果，X と Y に見かけ上の相関（擬相関，spurious correlation）が生じていると主張したわけです．

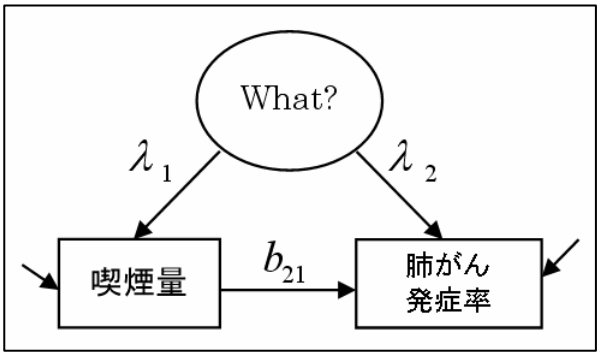


図２

3.2 婚姻率が上がると死亡率が下がる？

図３は都道府県別の婚姻率と死亡率（人口千人あたり）を散布図にまとめたものです．婚姻率が上がれば死亡率が減少するという負の相関関係が見て取れます．この関係は皆さんの直感に合うでしょうか．婚姻率と死亡率は当該年度に役所に提出された婚姻届と死亡届の数に基づいています．この散布図の結果を理解するには婚姻率と死亡率以外の変数--- 第三変数---を考える必要があります．

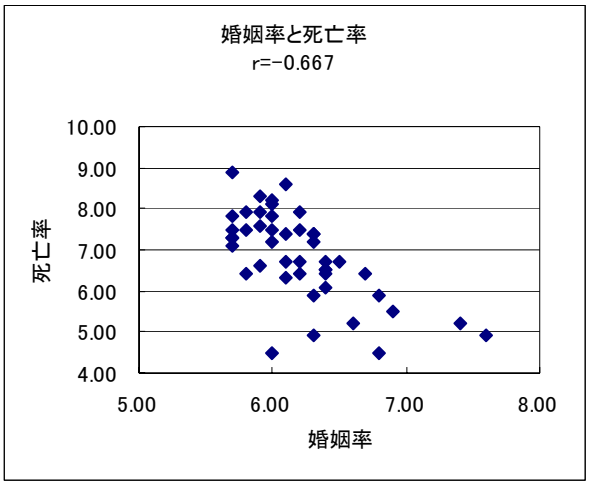


図３

3.3 子のつく名前の女の子は頭がいい

このタイトルの書物が出版されて約 10 年が経ちます．この主張について受講生の皆さんはどのように感じられるでしょうか．最近は多くありませんが，名付け親を頼んだり，漢字の画数にこだわったりすることがあります．命名という行為とその子の人生にはどのような関係があるのでしょうか．

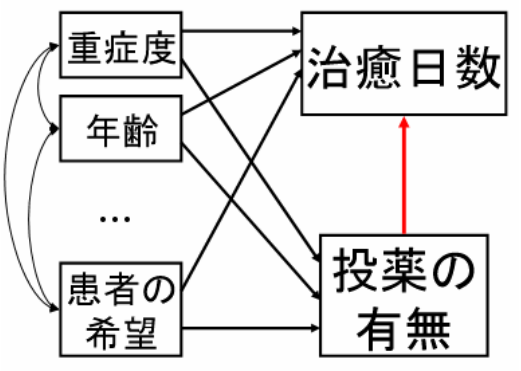
常識的には「関係がない」と考える人が多いように思いますが，もしそうならば，

なぜ「子のつく名前の女の子は頭がいい」という主張が登場するのでしょうか．ここでも先に考えた二つの例のように第三変数に着目します．

４．統計学的因果推論

この節では統計学的に因果推論がどのように定義されるか紹介しましょう．例として薬の服用が病気の治癒に効果があるのかどうかを調べたいとします．統計学的には，病気に罹った実験協力者を無作為に二つのグループに分け，一つ目のグループには薬を服用，二つ目のグループには服用せず，グループ間で治癒日数や治癒率を比較することが行われます．これを無作為割付け(random assignment)による実験研究と言います．各実験協力者をグループに無作為に割り当てることで，協力者のさまざまな特質が確率的にバランス化し，グループ間で公平な比較が可能になります．しかし，この方法は倫理的な問題をはらんでいることに気づくでしょう．たとえば，医者は重症の患者には投薬を勧めたいでしょうし，また，薬嫌いの（比較的軽い症状の）患者に無理やり投薬することは難しいからです．

そこで，医者と患者の合意で薬を服用したグループと服用しなかったグループにおいて投薬効果を比較するという方法が考えられます．これを観察研究（または相関研究）と言います．無作為割付けをした実験研究と比べると，二つのグループの非均質性に気づきます．たとえば，投薬グループは比較的症状の重い患者が，非投薬グループには軽症の患者が集まっていると考えられます．また，年齢による偏りがあるかもしれません．均質でないグループを比較しても投薬効果を適正に評価できるとは思えません．そこで次のように考えます．薬を服用した患者がもし服用しなかったとしたら治癒しただろうかしなかっただろうか．治癒したとしたら治癒日数は薬を服用した場合と比べて長引いたかどうか．薬を服用した患者に対して「服用しなかったら」と考えるわけですから，反事実(counterfactual)とか仮定法過去のモデルと言われています．



	投薬を選択しなかった患者			投薬を選択した患者		
	1	$\cdots$	m	$m+1$	$\cdots$	n
Y_0	Y_{01}	$\cdots$	Y_{0m}	$Y_{0,m+1}$	$\cdots$	Y_{0n}
Y_1	Y_{11}	$\cdots$	Y_{1m}	$Y_{1,m+1}$	$\cdots$	Y_{1n}
X	0	$\cdots$	0	1	$\cdots$	1
$\mathbf{z}$	$\mathbf{z}_1$	$\cdots$	$\mathbf{z}_m$	$\mathbf{z}_{m+1}$	$\cdots$	$\mathbf{z}_n$

図 4

各患者において，薬を服用しなかったときの結果を Y_0，服用したときの結果を Y_1 としますと図４(右)のようなデータが得られることになります．実際は，服用したかしなかったかのどちらかしかデータがありませんから，一方は欠測値(missing value)となり観測できません．表で四角で囲ってある部分が欠測値ということになってしまいます．この表では，投薬の選択と治癒日数に関係すると思われる第三変数，たとえば，「重症度」「年齢」「患者の希望」などを記号 z で表しています．このアプローチでは，適当な条件の下で，z の情報を活かして欠測値を推定し，投薬の効果（の期待値）$E[Y_1]-E[Y_0]$を推定します．

５．おわりに

統計学という学問は，遺伝学者でもあった R. A. Fisher によってその基礎固めがなされたと言われており，約１００年の歴史をもっています．近年はデータ科学とかデータサイエンスと呼ばれることがあります．日本統計学会は統計学を研究・普及する組織で会員数は現在約１５００人です．米国統計学会が約２万人，英国統計学会が約５千人の会員を擁することを考えますと，日本の統計学会の規模は小さめです．しかし，今回お話したデータの見方や統計的思考実験は研究活動だけでなく普段の生活においても重要だと思います．日本人はロジカルな会話と行動に不得手であるとよく言われますが，一方，数学の力は世界でトップクラスです．幼少から鍛えた算数・数学のスキルが現実の生活に活かし切っていないのではないでしょうか．統計科学の基礎は，数学と現実をつなぐインタフェースと考えることができますから，現代人の教養---読み書きそろばん---に加えられるべきだと思います．

参考文献

金原克範(1995,2001).『子のつく名前の女の子は頭がいい』洋泉社.

サルツブルグ(2001). 『統計学を拓いた異才たち』竹内・熊谷訳(2006). 日本経済新聞社.

狩野　裕(2002).「構造方程式モデリング，因果推論，そして非正規性」 竹内啓（編著）多変量解析の展開 -- 隠れた構造と因果を推理する -- pp.65-129 (Part II). 岩波書店.

東京大学教養学部統計学教室編(1994). 『人文・社会科学の統計学』東京大学出版会.