オリエンテーション
～統計的因果推論～

情報・システム研究機構理事

統計数理研究所長

椿　広計

概念とその関係性:用語学（論理学）

ISO 1087-1:2000 Terminology work—Vocabulary－Theory and Application

- 概念（Concept）＝対象の集合（外延：Extension）
  - 必要十分な特性（内包：Intension）の提示で概念は定義される：
    - テロ事象＝暴力性＋政治性＋直接被害者を超えた影響
      - CIA (1981)Patterns of International Terrorism
- 概念間の関係性
  - 一般化(Generic Relation)：上位概念＋追加特性＝下位概念
    - マネジメント⇒品質マネジメント（品質に対する＋マネジメント）
      - 統計的測定モデル（テロか否かといった論理問題からテロらしさの測定）
  - 部分（Partitive Relation）：ある概念の一部となる
    - 四季＝春夏秋冬⇒統計的分類問題の必要性
  - 関連性（Associative Relation, Pragmatic Relation）
    - 経験（データ）が繋ぐ非階層的な概念の繋がり：統計的関連性の追求
      - 逐次（Sequential）関係：生産⇒消費
      - 因果（Causal）関係：アクション⇒反応

夏目漱石：文学評論（岩波文庫）、初版1909
1903の東京大学文学部での記述統計学的講義

- その道の人は科学をこう解釈する。
  - ロンドン留学中K. Pearson 「科学の文法」第2版勉強の成果
- 科学はいかにしてということすなわちHowということを研究するもので、なにゆえということすなわちWhyということの質問には応じかねるというのである。
- たとえばここに花が落ちて実を結ぶという現象があるとすると、科学はこの問題に対して、いかなるプロセスで花が落ちてまたいかなるプロセスで実を結ぶかという手続きを一々記述してゆく。
  - Sequential Relation
- しかしなにゆえに花が落ちて実を結ぶかという問題は棄てて顧みないのである。（中略）

Ａｎａｌｙｓｉｓ：分解

- さてこのいかにしてすなわちHowということを解釈すると、俗にいう原因結果という答えが出てくる。
- しかしまえに述べたようなわけだからこの原因結果とは**ある現象の前には必ずある現象があり、またある現象の後には必ずある現象が従うという意味**で、甲が乙をしかならしめたなどという意味ではないのは無論である。
- それでこの原因結果を探るには分解をする。
- 一つの現象をとって『いかにして』ということを究めるには、それが複雑な現象であればあるほど『いかにして』ということを知りにくい。
- 知ったと思うても分解を経た上でないと常に間違う。だから**人間はその場合とその時代に応じてでき得るかぎりの分解を企てる**。

Ｓｙｎｔｈｅｓｉｓ：総合

- 分解をしてある微細なことについて
  『いかにして』ということが分かると、
  つぎにはこの零細なる事実をたくさん集めて比較してみる。
- そこで総合ということが始まる。
- 総合とは同じような事実をたくさん集めて
  『いかにして』という点において
  みな一致していることを見ることである。
- で総合ができれば、これから一つの法則ができるわけである。

分類：Classification

- それから総合をしてみて『いかにして』という点において
  いろいろな場合が一致しなければ分類ということができる。
- まず**ざっとこんなふうで科学はできる**。
  - 留学中の漱石文学論ノートより一部椿現代語訳抽出
    - **観察や実験で得られたデータを分析し、抽象化し、さらに総合し、一般化し、分類することで法則ができる。これが科学である。**
    - **科学が分類や一般化によって法則を導くのは、実用的便宜性からである。**
    - **実用的便宜性とは、過去から未来を予測できることである。**
    - **昔から、発明家は想像から示唆を得て、その後に検証を行った。**
    - **検証が難しくても、この仮説的法則が成立する確率が高ければ、その確率に比例して法則は有効である 。**

関連性としての因果関係

- メカニズムに迫る統計モデリング
  - 有向関連性と因果関係（法則性）
    - ＸＸという環境では
      - Aを原因としてBが起きる：A→B　この薬を飲むと病状が改善する
      - Bを原因としてAが起きる：B→A
      - どちらも原因となりえる：A ⇔ B：両方向因果
        - 経済政策を行うと景気が回復する And 景気が悪いから経済政策を行う
- 記述的統計モデリング（メカニズムに迫る前段階でもちろん有用）
  - 無向関連性と相関関係
    - AとBとには関連性がある：A－B　語学ができると数学もできる
      - 潜在変数（学力因子）による有向関連性探索への発展可能性
    - 偽相関発見を経緯とした原因系変数探索への発展可能性
      - 年収が下がると短距離走の成績が上がる→年齢を調整すれば消える相関

統計的方法と因果推論

- 「実験」による実証
  - 原因Aを介入的に変動させ、結果Bを観測する
  - **Fisherの「実験計画法（Design Of Experiments）」**
    - 無作為化実験の原則（対象に原因をランダムに割り付け）と繰り返し実験の原則
      - さらに実験から得られる情報量を最大化する様々な数理：直交計画等
  - 実験的方法にかかるコストと人や動物を対象とする場合の倫理制約
- 「観察」による代替可能性
  - 因果仮説に関する情報が十分で定量性だけ知りたい場合
    - 確証的方法論：パス解析→SEM (Structural Equation Modelling)
      - パス解析（因果構造モデル）と因子分析（測定モデル）の統合モデリング
        - 行動計量学分野で発展
  - **今日のテーマ：因果関係に関する情報が不十分で検討の必要性**
    - **介入実験や介入後のフォロウアップ（追跡）調査の代替手段（模擬実験的方法）**

無作為化のマジック
観測された世界とされなかった世界の按分

- 2個の錘があります。錘１＝Xg, 錘2=Yg でX,Yは途です
- 秤を1回だけ使って，2個のの重さを偏りなく推定(Estimate)してください

■回答：適当に重さを適当に推定します.

- 錘1：100ｇ，錘2: 100g：多分間違っています

■コインを投げます

- 表が出たら錘１を計ります。裏が出たら錘２を計ります：与えられた確率でデータを観測

■次のように重さを報告します.

- 表がでたら⇒錘１の重さの推定値＝2X-100, 錘２の重さの推定値=100⇒合計は$2X$
- 裏が出たら⇒錘１の重さの推定値＝100, 錘２の重さの推定値＝$2Y$-100⇒合計は$2Y$
  - *錘1の推定量の期待値＝（2X-100 +100)/2=X,　錘2の推定量の期待値=Y*

■この原理を上手く使うには⇒調査対象集団（母集団）を正確に定義

■そこから確率的に対象を観測する（選択する）：Neymanの標本調査論

データ解析に影響を与える観測されなかったデータ

■多くの教科書的データ解析ではデータの完全性や等質性が前提

✓実際のデータ解析: 完全性の欠如，異質性の混入

■問題を引き起こすデータ

✓**観測の不完全性**(Incompleteness)

- **欠測値(Missing　Data)**:
  - **単位欠測（丸ごと欠測：Unit Nonresponse）**：回答者が全ての項目に非回答：調査票が戻ってこない
  - **項目欠測(Item Nonresponse)**：ある項目に回答せず：この項目には回答しない
- 計測の不完全性：測定の誤差（Measurement Error)
  - 系統的誤差(Systematic Error)
    誤ったデータが報告され，照会によりデータを修正可能
  - 偶然誤差(Random Error)
    観測精度(Measurement Accuracy)起因：丸め誤差：確率的誤差として処理

✓観測の異質性

- 外れ値(Outlier)：他のデータとは異なる挙動を示し分析に影響を与えるデータ
- 影響力のあるデータ(Influential Data):データ分析結果に影響力のあるデータ

Little and Rubin (1987) *Statistical Analysis with Missing Data*, Wiley. の欠測（未観測）のメカニズムと分類

■比較的単純処理可能な欠測(Ignorable Missing)

⇒ランダムな観測，ランダムな欠測（未観測）

➢データYが観測される確率 $0<p_Y$ が想定可能，欠測確率 $q_Y=1-p_Y$

● 完全にランダムな欠測（MCAR:Missing Completely at Random)

欠測確率が一定：典型例：完全無作為標本（**抽出率＝観測確率**）

● **ランダムな欠測 (MAR: Missing at Random)**

- **観測確率が他の変数で表現可能：** $p_Y=f(x_1,\ldots,x_p)$

  Rosenbaum and Rubin (1983) The central role of the propensity score in observational studies for causal effects, *Biometrika*, 70, pp.41-45.

- 典型例：層別無作為標本（層毎に抽出率【観測確率】が変わる）

■**無視できない欠測**： (Non-ignorable Missing)：実データの大半はこれ

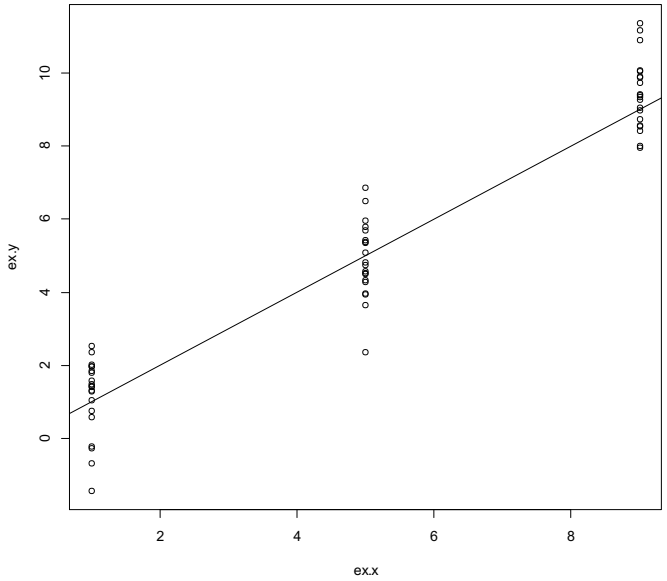
➢ただ事では済まない欠測

➢欠測確率 q_Y が欠測して観測できないデータYに依存」こんな値出したくない

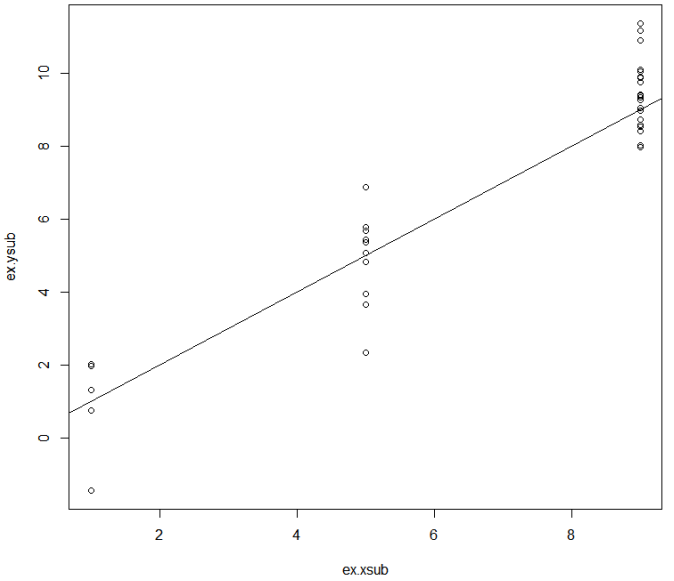
MARの状況での目的変数欠測
回帰への影響は何故無いのか？

完全データ $Y=X+\varepsilon$，　ε ：標準正規乱数
X=1,5,9に Yがそれぞれ20個分布

目的変数 Yの欠測確率が
説明変数Xで決まる場合



実線：真の回帰直線　Y=X



X=1⇒25%，X=5⇒50%，X=9⇒100%観測

仮想世界のシナリオ

■全国1892市区町村：実世界全体

- ➢解析除外
  - ● 人口0の4町村（福島第一発電所周辺）
  - ● 政令市は除き区データを利用

■分析に利用するのは4変数(2015年)

- ■（独）統計センター教育用標準データセット(SSDSE)
- ➢人口総数，高齢人口，世帯数，出生数

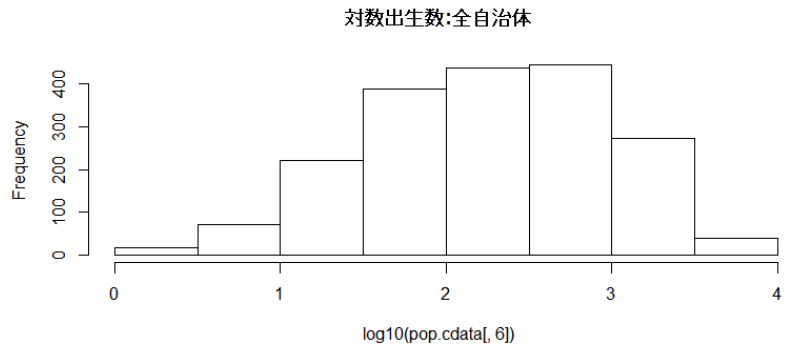
■仮想的欠測（観測）シナリオ

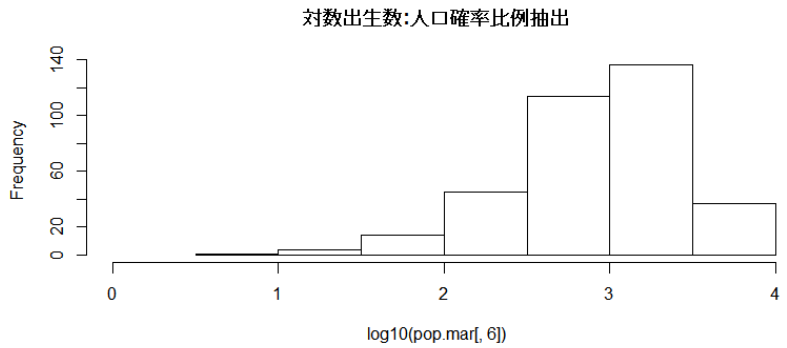
- ➢人口総数，高齢人口，世帯数は全自治体で観測
- **➢出生数はある自治体群でのみ観測**

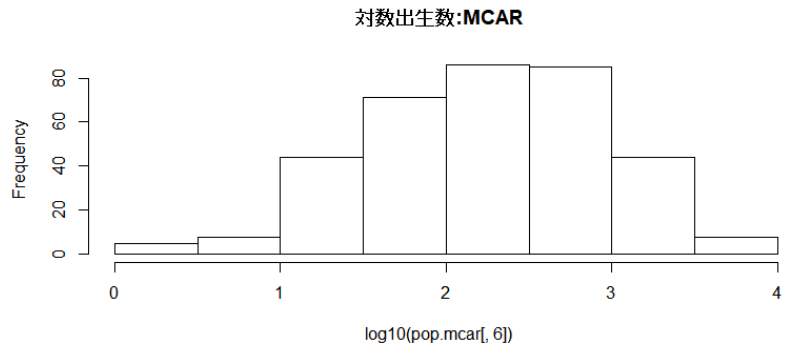
■目的：全国市区町村の平均出生数を推定したい

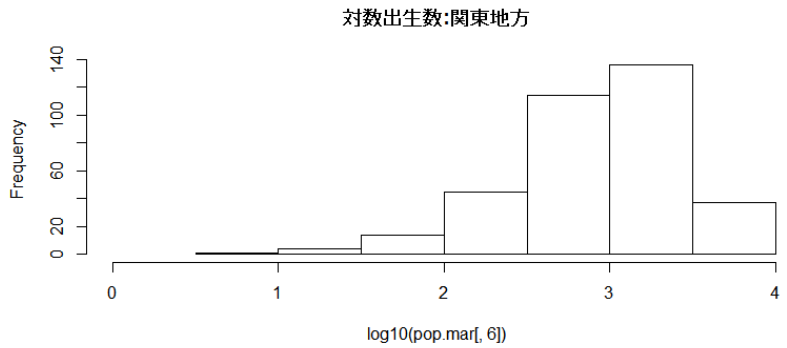
観測データの対数出生数のヒストグラム

MCAR（完全無作為抽出351自治体）MAR(確率比例抽出：351自治体)・関東地方351自治体の分布









数理統計学の2つの背反する主張
反事実(Counter Factual)を巡って

■ Conditional Inference vs. Unconditional Inference

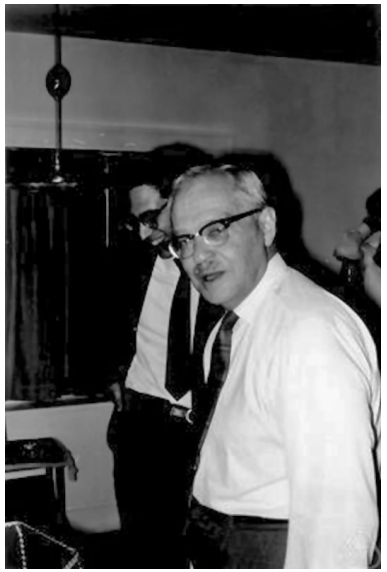
- ✓ Fisher流**条件付推論**：起きてしまったことは確率変数とは考えず所与として対処
  - 欠測値が観測された可能性は配慮すべきでない
- ✓ Neyman流**無条件推論**：起きなかった可能性（**反事実**）を配慮して対処
  - 欠測値が観測された可能性を配慮

■ 欠測値解析はこの両接近で異なる実務が生じる問題

■ 標本調査論はNeymanの流儀

- ✓ Rubinの**Propensity Score(観測確率推定)**法

https://ja.wikipedia.org/wiki/ロナルド・フィッシャー



https://en.wikipedia.org/wiki/Jerzy_Neyman#/media/File:Jerzy_Neyman2.jpg

Fisher流の集計???

観察されていることを前提の**条件付推定（観測値を用いた回帰分析等）で**
欠測値を補完(Impute)し、欠測値は補完値に置換して集計(Conditional Estimation)

欠測値推定値と実測値とで集計

- ×関東データの出生数の要約統計量

```
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
    1.0   157.5   532.0   960.6  1316.0  8019.0
```

- △関東もそれ以外も欠測値推定値（回帰分析）を用いた場合の要約統計量

```
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
   0.00   44.89  172.16  492.82  596.11  8026.09
```

- ○関東は実測値，その他は欠測値推定値による集計

公的統計実務でやっていること

```
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
   0.00   44.94  172.29  492.82  599.15  8019.00
```

- 全自治体の正解

```
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
    0.0    54.0   199.0   531.3   656.0  8019.0
```

欠測に対するNeyman流の偏りの無い補正

■MARの場合≒層化無作為標本

➢観測確率p_Yあるいはその推定値$\hat{p}_Y$を用いた観測データを含む
全データの補完ではなく補正Y_C

- **欠測値補完：Yが欠測ならば，補完値$\hat{Y}$を作成し，これを欠測値に補完（Impute)**

  $Y_C = Y$

- **観測値修正：観測値Yも欠測していた可能性(確率：1- p_Y)を配慮して次の値Y_Cに補正**
  - 古典的には「差推定」と呼ばれる：公的統計実務では乗率を掛けることに相当
  - $Y_C = Y + (Y - Y) / p_Y$

➢反事実(Counter Factual): Yを観測した世界と観測しなかった世界

- 補正値Y_Cの期待値はY：　観測確率がわかれば**補完値の偏りが補正**される
- 証明：$E[Y_C] = p_Y \times \{\hat{Y} + (Y - \hat{Y}) / p_Y\} + (1 - p_Y)\,\hat{Y} \quad = Y$

■MCARの場合：MARの特別な場合⇒完全無作為標本
状態推定（母平均，母標準偏差）も構造推定（回帰係数など）も欠測を無視可能

■無視できない欠測⇒解析に決め手はない！
MARの世界が近似的に成立していると見なして不完全ながら補正

FisherとNeymanの仮想世界での喧嘩 実世界でも仲が悪かった

■Neymanの逆襲：Fisher流の集計には偏り

- 欠測推定値$\hat{Y}$が使われる確率は$1\text{-}p$
- 観測値Yが使われる確率はp
- その期待値は，$pY+(1\text{-}p)\,\hat{Y}$
- $\hat{Y}$がYの偏りのない推定量でない限り，期待値はYにはならない

■Fisher:現実に生じたYを確率変数として扱うのは疑問

- Yを定数で所与として統計的推論は行うべき
- 偏りを除去したNeyman流補正は集計の予測平均二乗誤差を増大させる
- 仮想現実を入れるべき状況とそうでない状況がある

統計的因果推論を使いこなす面白さに触れてください

それでは本論へ