探索的データ分析

労働経済学 2

川田恵介

- 数理分析 (データ分析 & 経済モデル) は有益な分析ツール
  - 致命的な落とし穴を避けることができる
  - ただし前知識が必要
  - 特に” 探索的” データ分析において、今後深刻化する恐れ

探索的データ分析

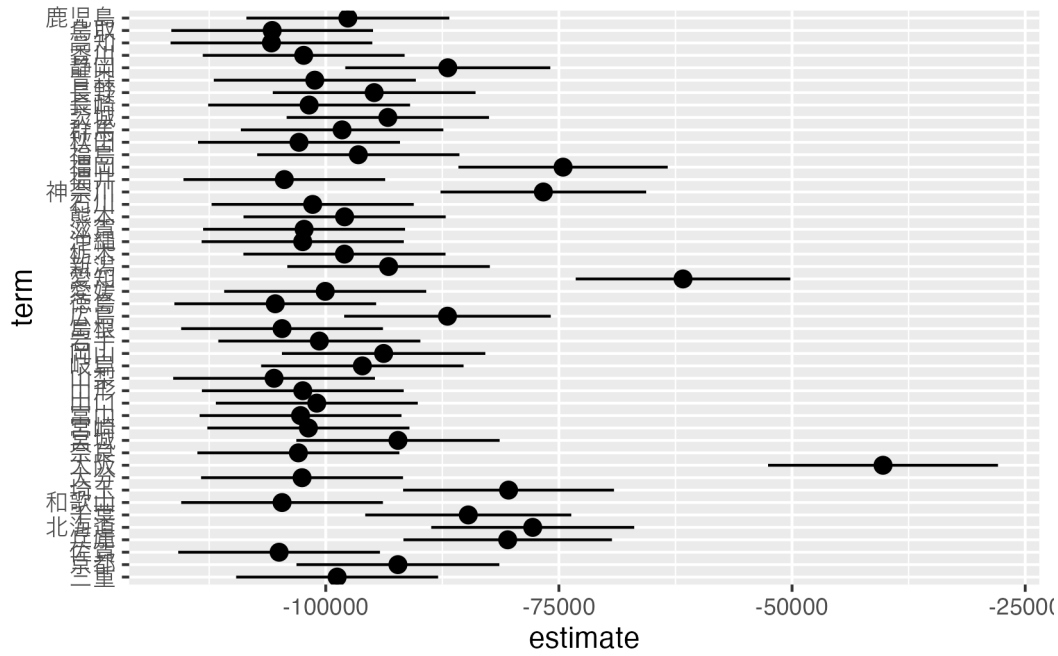
- Y と変数群 X の関係性をざっくり把握する
  - 線形近似モデルを用いる
  - 関係性の原因は問わない
- 因果推論: 因果的な関係性を把握する
- 経済モデル: いくつかの基本アイディア（インセンティブ、資源制約等々）を元に関係性を整理する
  - 比較的少数の変数に焦点を当てる
- 労働分析において有益

探索的線形モデル

$$E[Y|X_1,..,X_L] = \underbrace{\beta_0 + \beta_1 X_1 + .. + \beta_L X_L}_{\text{全て関心}}$$

- Mincer 型賃金モデル: $X =$ 年齢、性別、学歴、勤続年数
- 都道府県間賃金格差: $X = 47$ 都道府県

例



多重検定

- 複数のパラメータを推定したい
  - 賃金と性別と学歴の関係性をざっくり比較したい
  - 各都道府県間格差を推定したい

復習: 信頼区間

- 信頼区間: “同じデータ活用・サンプリング法を用いる研究者” の $1-\alpha$ が正しい値を含む信頼区間を獲得できる
  - “間違った” 区間をえる研究者割合 α をコントロール
- 特定の推定値 (Point-wise) について Valid な信頼区間
  - 大量の推定値について適用できるか？

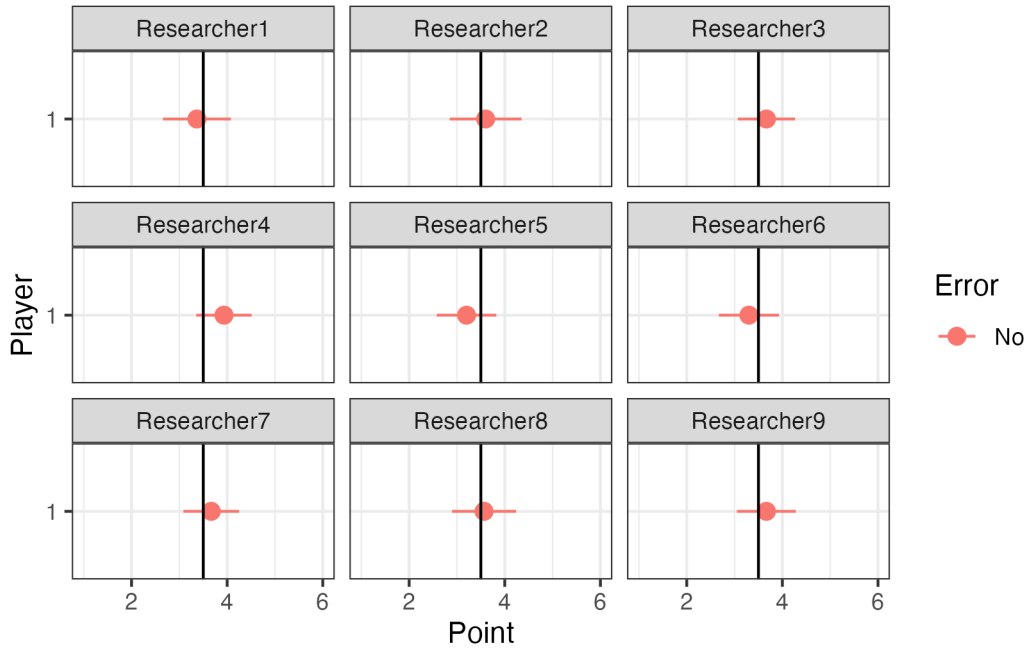
例

- Research question:「サイコロの出目をコントロールできるか？」

  - サイコロの出目の平均値を操作できるか？

- 一人のプレイヤーのみを収集するのであれば、Defalt は有効

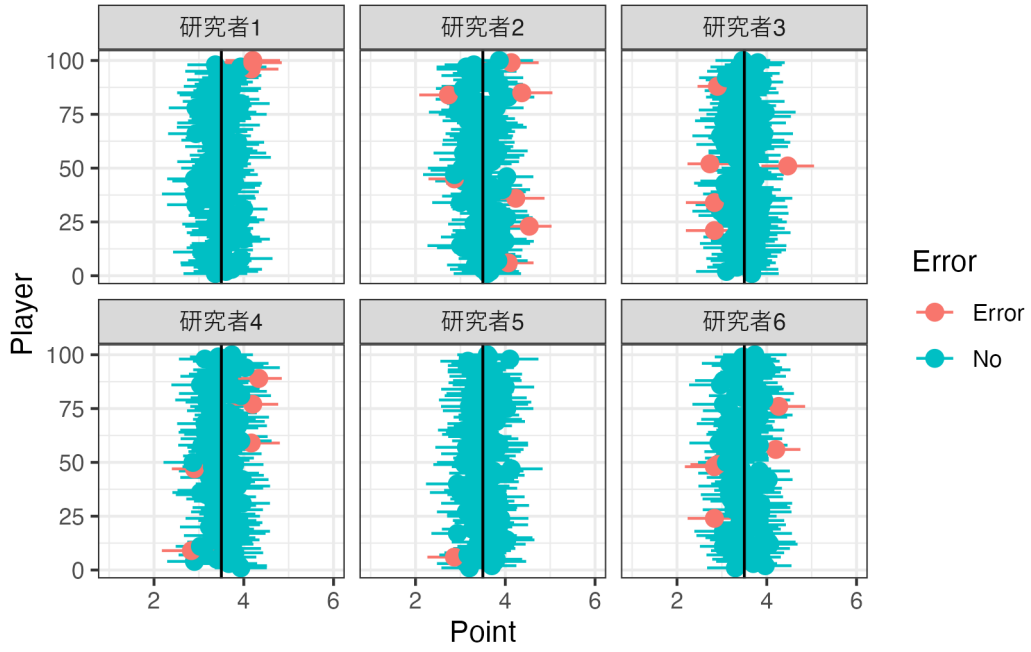
適切: 1 player



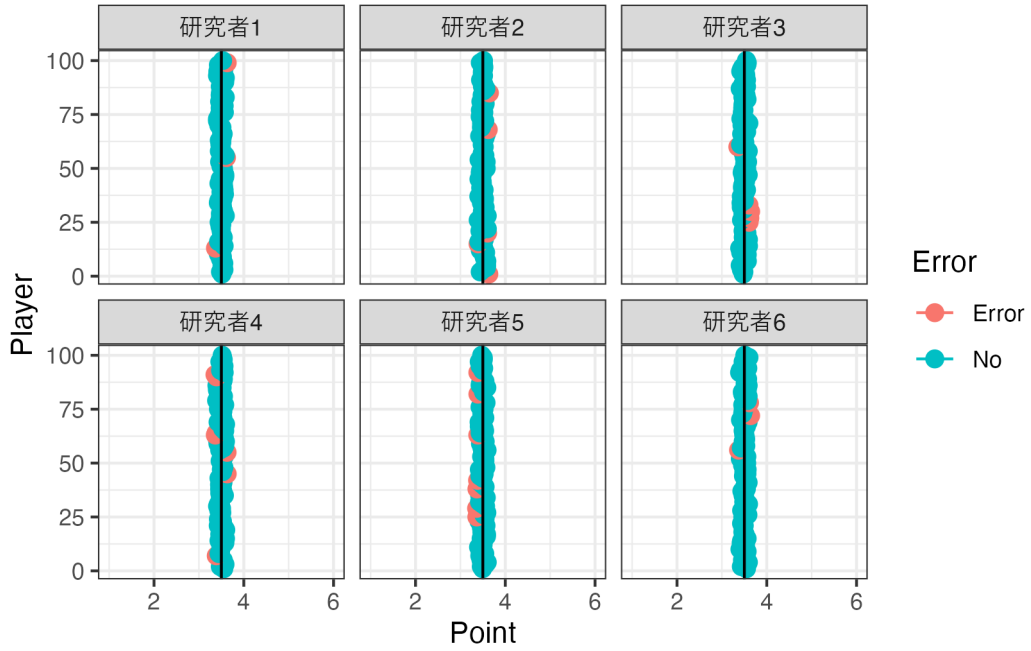
例

- Research question:「サイコロの出目をコントロールできるか？」
  - サイコロの出目の平均値を操作できるか？
- 「100 名のプレイヤーのなかから、操作できるプレイヤーを見つけられるか？」であれば、不適切

不適切: Small sample



不適切: Larger sample



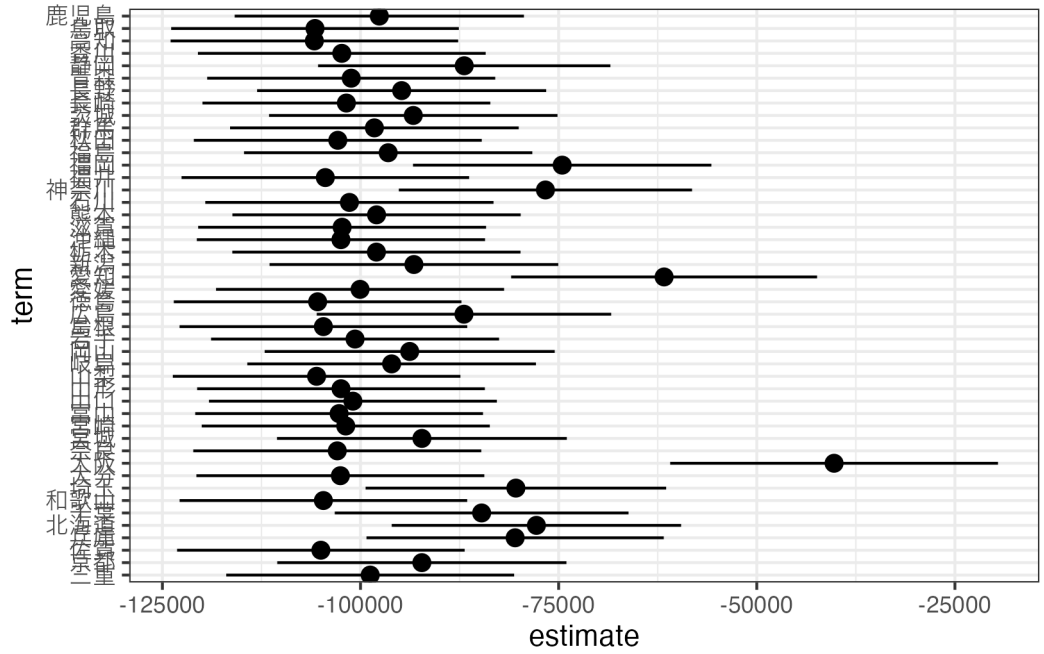
問題点

- 「本来はほとんど差がないのに、サンプルが偶然偏った結果、大きな差が推定される」
  - なんとしてでも注目されたい研究者にとっての、「当たりくじ」
- 「どんなに当たり確率が低いくじであったとしても、無限回引けば” 絶対” に当たる」
- 「ある一人のプレイヤーについて、間違えを犯す研究者の割合」≤「１００名中一人以上に間違えを犯す研究者の割合」
  - 一般的に生じる問題
  - BigData は解決しない

Family-wise confidence intervals

- Family-wise confidence interval: 複数の推定値が前提
  - **一つ以上**の信頼区間が真の値を含まない確率を α_{Family} 以内に抑える
- Point-wise confidence interval: ある一つの推定値について、信頼区間が真の値を含まない確率を α_{Point} 以内に抑える
- Bonferroni 法: $\alpha_{Family} = \alpha_{Point}$/推定値数

例



発展: Bonferroni **法の根拠**

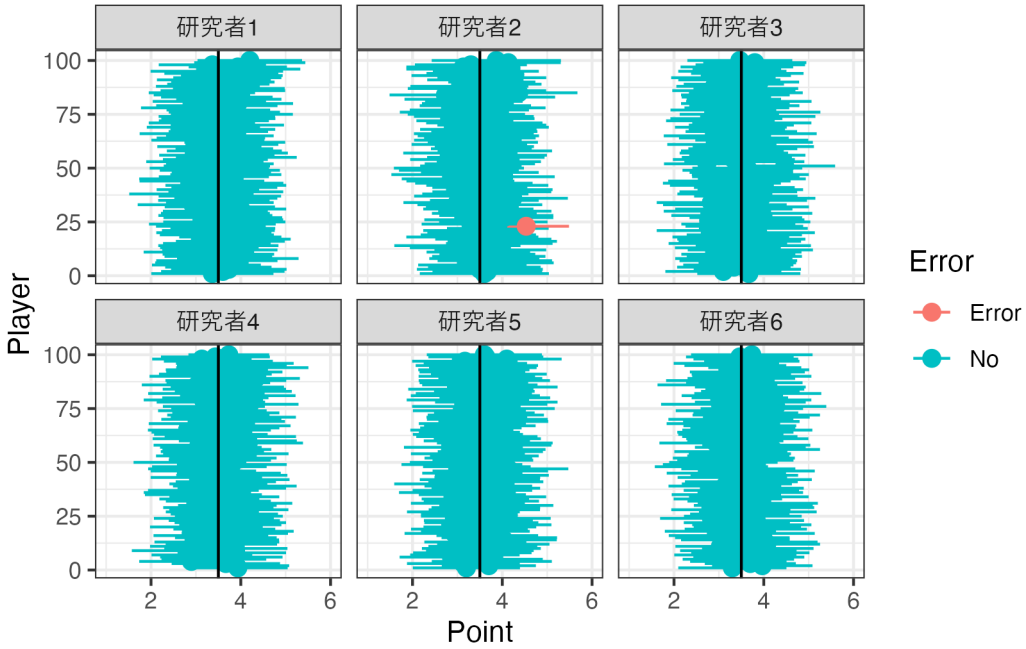
- 以下の一般原則を利用

一つ以上の区間について間違いが起こる確率 $(= \alpha_{Family})$

$\leq$ 区間1について間違いが起きる確率

$+ \ldots +$ 区間Lについて間違いが起きる確率 $(= L \times \alpha_{Point})$

適切: Larger sample with adjustment



発展: 他の手法

- 統計的検定を発見的に使いたい場合、0.5% に設定すべきという主張も
- 多重検定問題への対応法研究は進む
  - 検定したいパラメータが非常に多くなると、Power が大幅に悪化することが動機
- 改善例: False Discovery Rate, Uniform inference などなど

再定式化

- 線形モデル $E[Y|X_1, .., X_L] = \beta_0 + \beta_1 X_l + .. + \beta_L X_L$
- 一般に β_0 は解釈しにくい
  - $X = 0$ の場合の Y の平均値。。。？
- X を中央化 (Centralized) する
  - $\tilde{X} = X - mean(X)$
- β_0 = 全ての X が平均値であった場合の Y の平均値

まとめ

- なんの考慮・対処をせずに多重検定を行なっている事例は極めて多い
  - いくらでも超能力者を”発見できる”
- 発見的に統計分析は慎重に