モーメント法

機械学習

川田恵介 (keisukekawata@iss.u-tokyo.ac.jp)

Table of contents

1　シンプルなモデルの推定

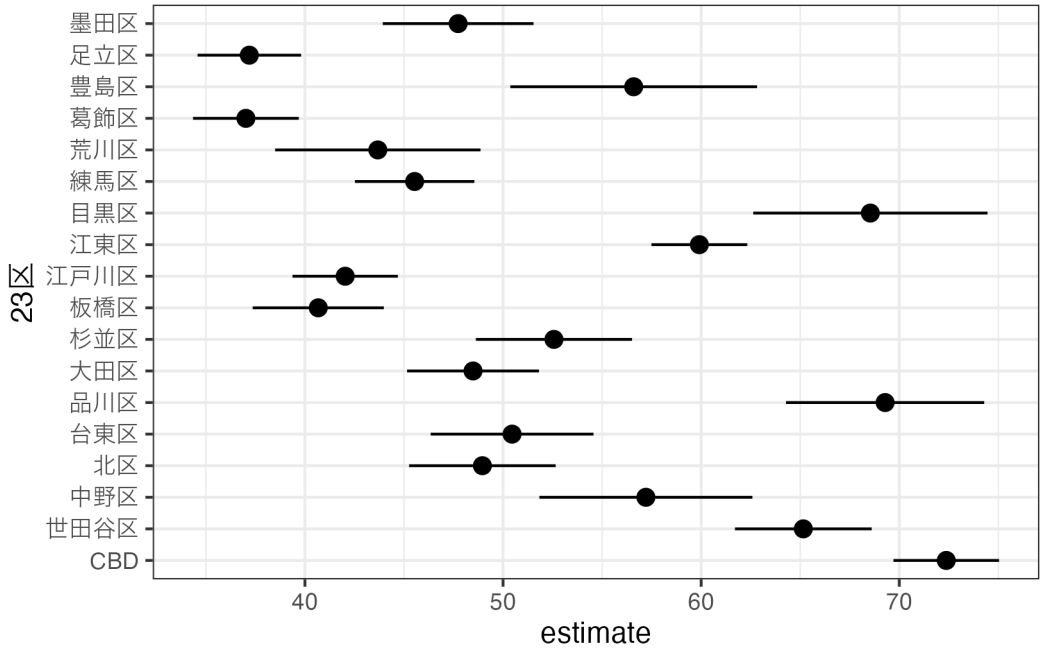
- 事例数に比べて、β の数が少ないモデルであれば、OLS や平均値は有力な推定方法
- OLS や平均値などを含む推定方法

1.1　それぞれの特徴

- 機械学習は、ある程度の事例数はあるが、X の組み合わせがそれ以上多い場合に有効
  - 平均値と近い値が推定できるように設計されているが、具体て k にどの程度近いのかは不透明
- 平均値や OLS は、“十分な事例数のもとで、平均値を知りたいのであれば”、が有力
  - 推定値の分布が明確であり、推定誤差 (信頼区間) を計算できるため

1.2　実例

- 信頼区間: 横線は、95% の確率で母平均を含む



1.3　分布

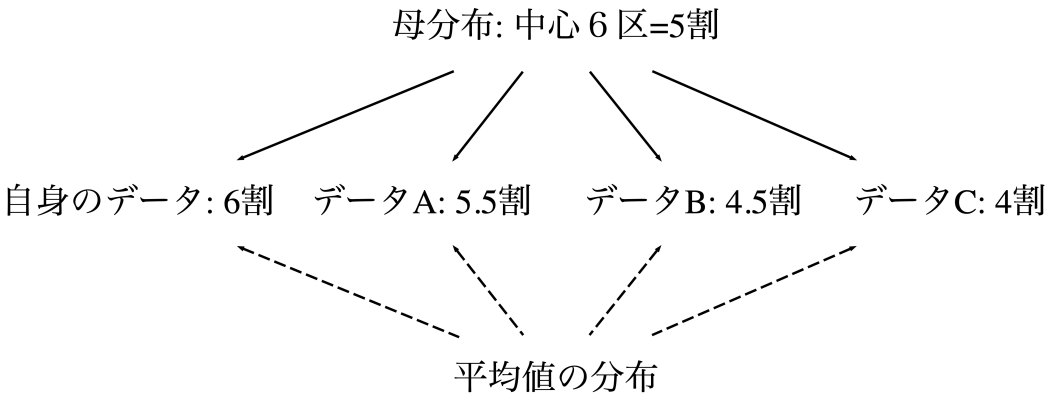
- 一般的な定義: 値のばらつき方

- 種類の分布を区別することが重要
  - 個別事例値のばらつき (事例の分布)
  - 平均値などの、事例の分布を特徴づける値のばらつき

1.4 事例のばらつき例

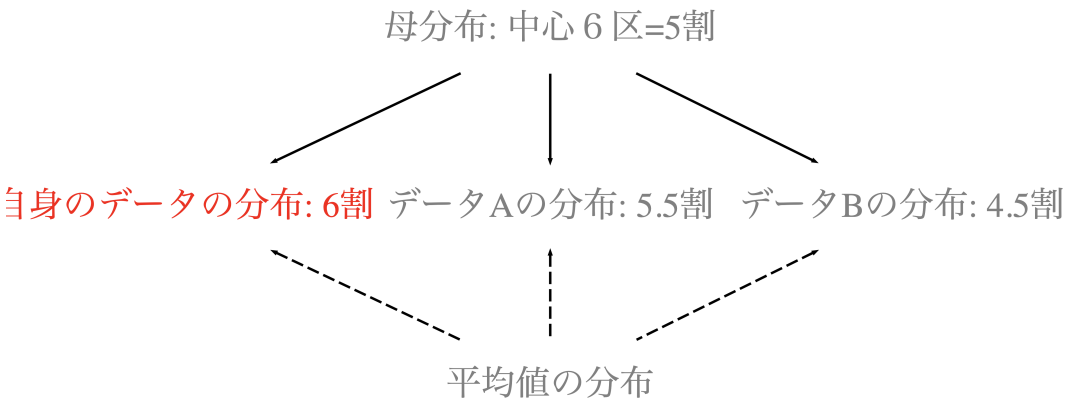
- 自身のデータにおける Y のばらつき
- 仮想的な母集団における Y のばらつき
  - 上記は一般に一致しない
  - 事例から計算される値 (平均値など) も一致しない

1.5 イメージ図



1.6 イメージ図

- 赤色: 実際にわかる/灰色: 実際にはわからない



1.7 データ分析の限界

- 母平均と推定値は一致しない
  - 推定値 = 母平均 と主張すると、“100%” を嘘を主張してしまう

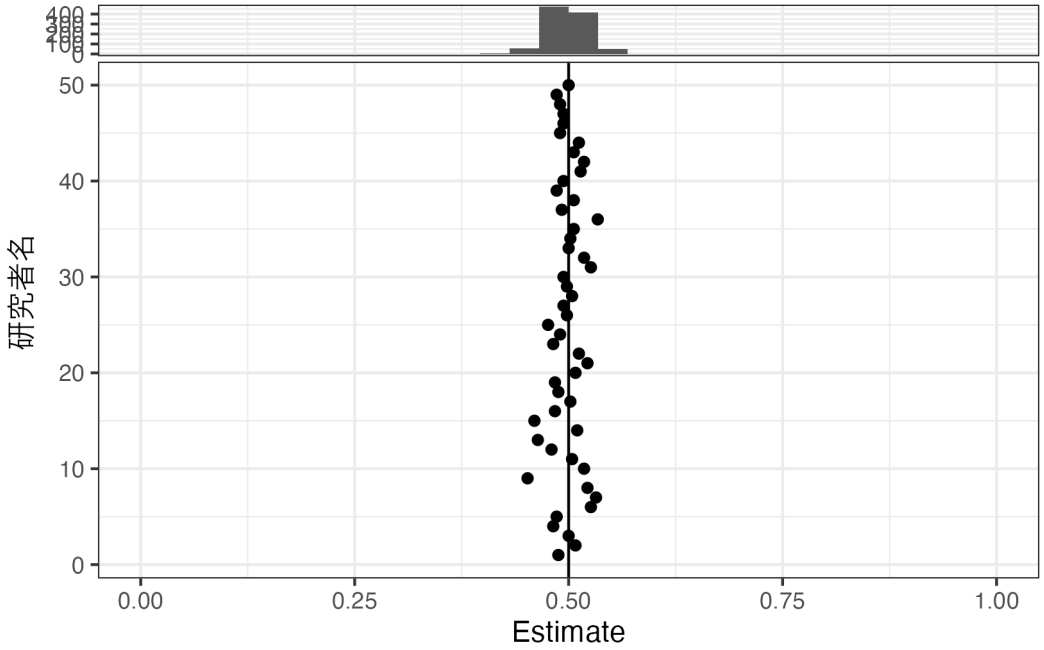
1.8 Yの平均値の利点

- Yの平均値の分布が、理論的に類推できる!!!
- 事例数が十分あり、ランダムサンプルングであれば、
  - 平均値の分布は正規分布 (ベル型分布) で近似でき、具体的に推定できる
    - 中心極限定理と呼ばれる
  - 数値実験より、200 事例以上を要求
- 平均値を保管する指標、信頼区間を計算できる

1.9 数値例

- Yの平均値を推定する
  - 母平均は 0.5
  - 事例数は 500
  - 1000 名の研究者が独自にデータを収集し、計算する

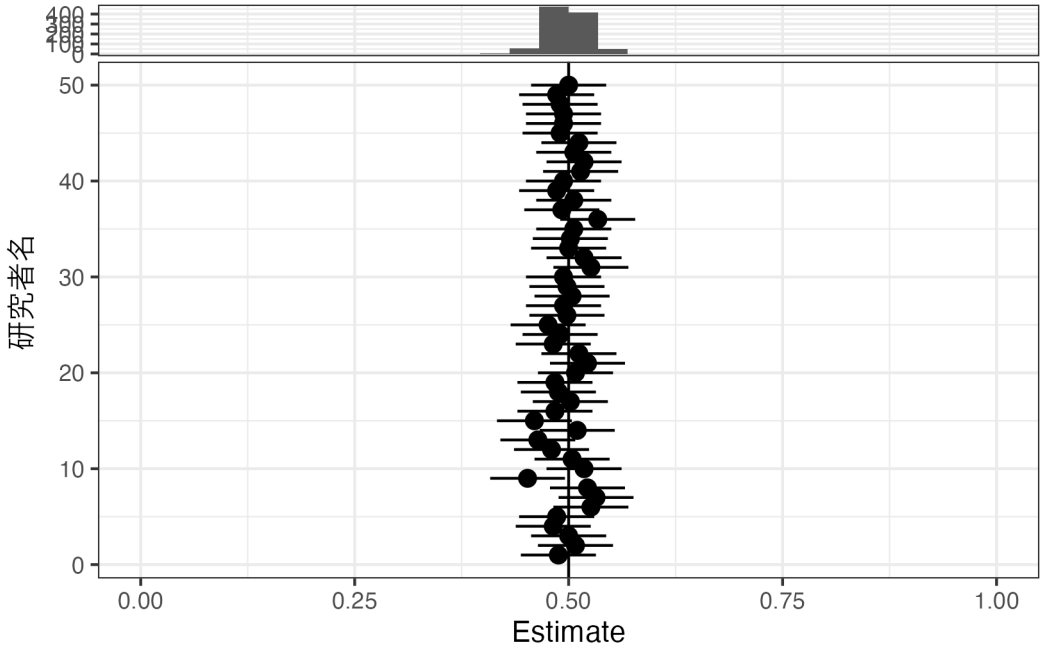
1.10 数値例: 平均値



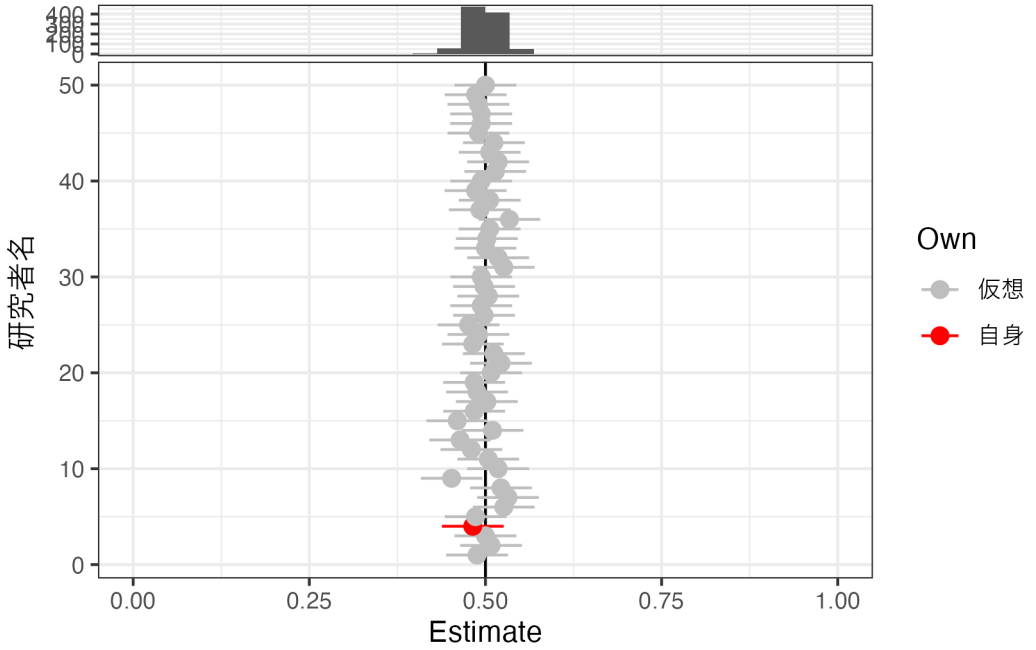
1.11 信頼区間

- データから計算される区間
  - 一定割合 (初期値では 95%) のデータについて、母平均を含む区間
  - 「データ上の平均値は正規分布に従う」、という性質から計算できる

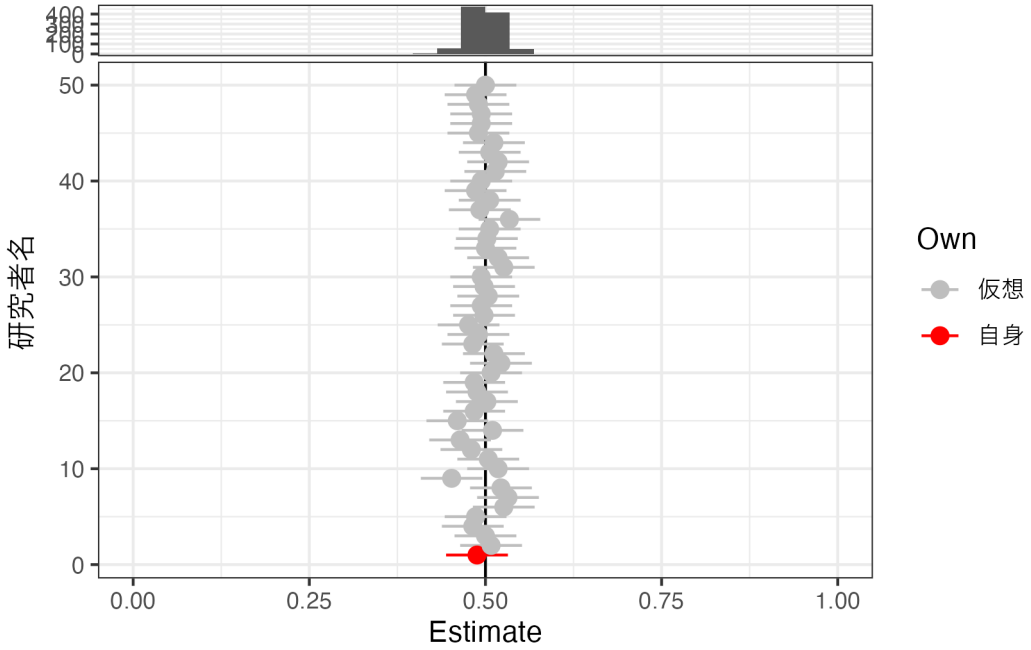
1.12 数値例: 信頼区間



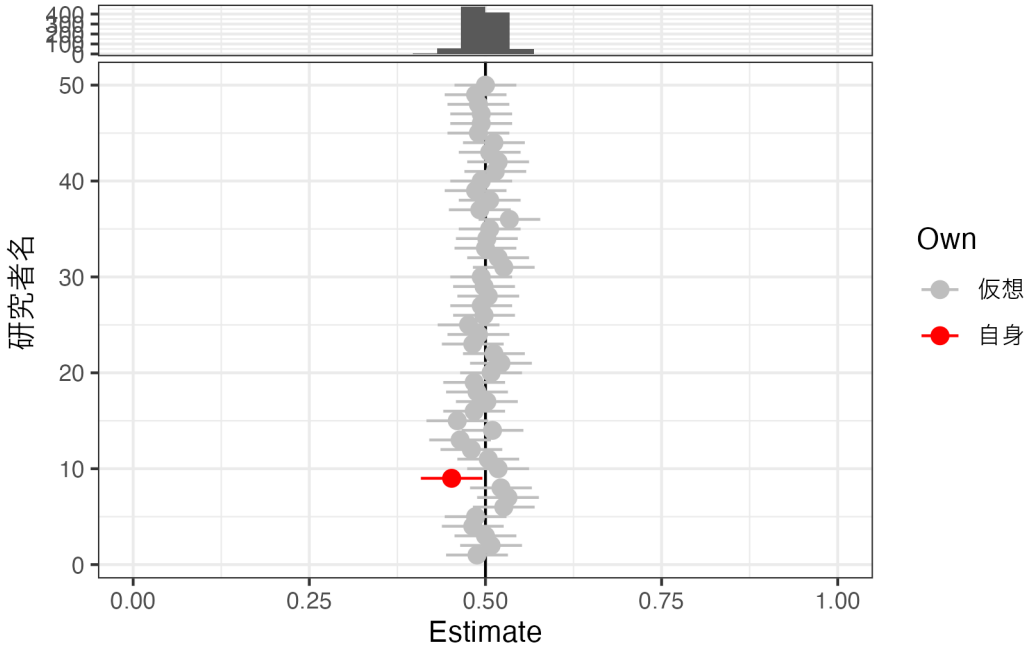
1.13 数値例: 現実



1.14 数値例: 現実



1.15 数値例: “不幸な” 現実



1.16 確率的意思決定

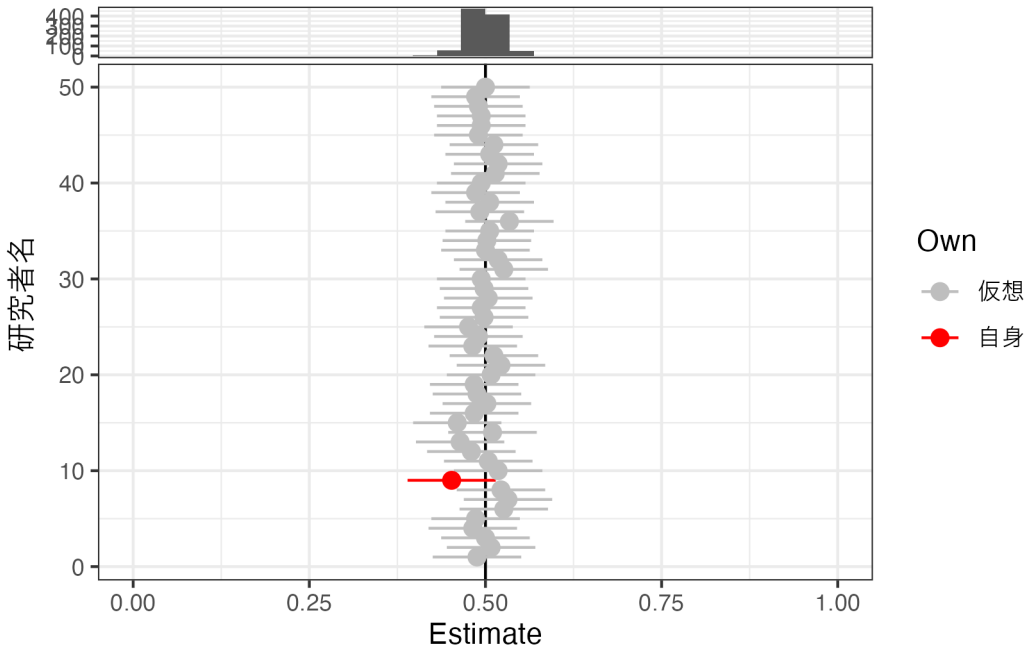
- 一定確率で必ず” 間違いを犯す” ことを許容する必要がある
  - 保険への加入、薬の服用等々
- “分析結果” も同様
  - 信頼区間を結果として報告する場合、一定確率で母平均を含む区間を報告してしまう
  - 初期設定では 5%
    - じゃんけんで 3 回負ける確率と同程度

1.17 性質: 信用確率

- 信用確率 (100 - ミスを犯す確率) %
  -「ある主張は信用確率が高いので採用する」
  - 仮想的な研究者の中で、母平均を含まない値を得る人の割合
    - 類似概念: 火災に遭う人の割合
- 信用確率を上げるためには、信頼区間を広げる必要がある
  - 100% 間違わない信頼区間を得るためには、無限大に大きな区間が必要
    - 分析結果を曖昧にしすぎており、“意味がない”

1.18 数値例: 現実

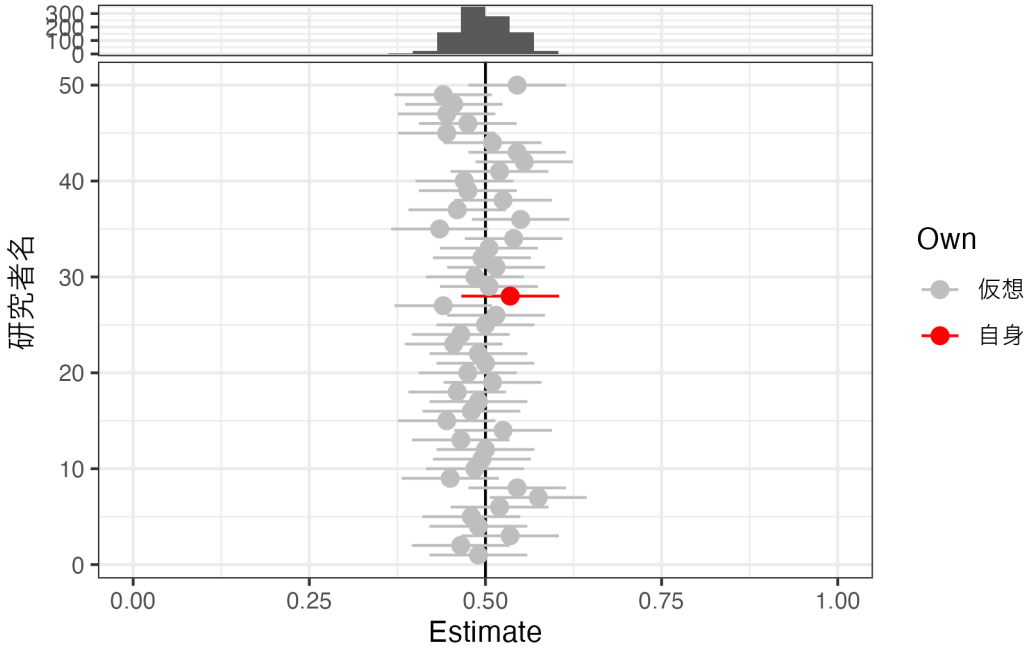
- 信用確率を 99.5% に変更 (じゃんけんで 5 回負ける確率と同程度)



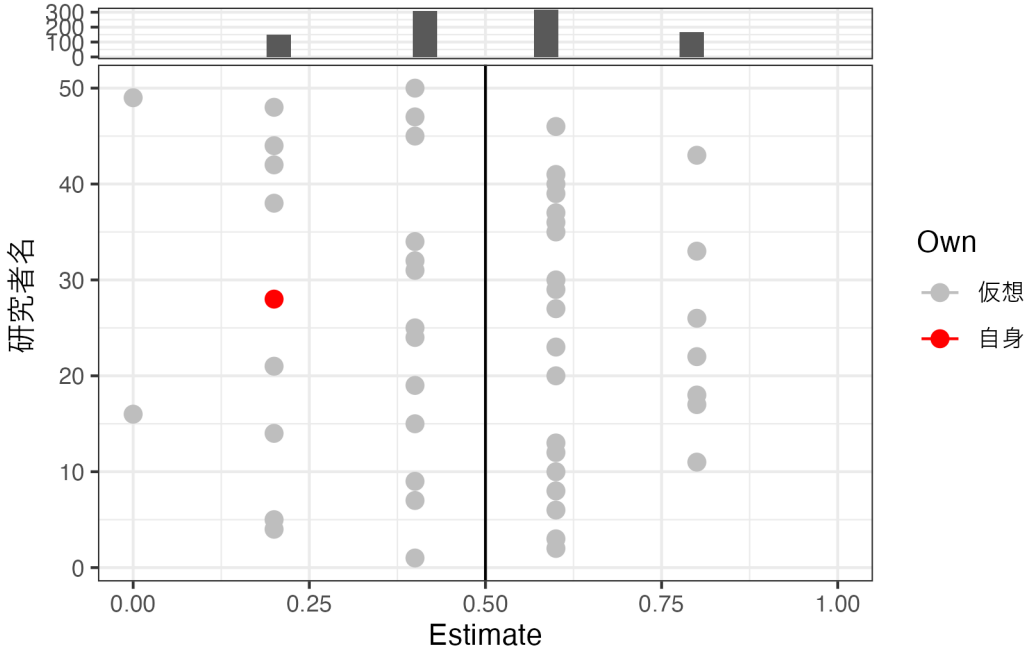
1.19 性質: 事例数

- 平均値を計算するために用いた事例数の減少は、
  - 信頼区間を広げ、結論を曖昧にする
- 大幅な減少 (100 事例を割るなど) は
  - 信頼区間の計算が不可能になる
    - 代替案: 最尤法/ベイズ法による推定
      - 問題点: より非現実的な仮定が必要なケースが多い

1.20　数値例: 小規模事例



1.21　数値例: 小規模事例



1.22 まとめ

- OLS や平均値は、β の数が事例数に比べて十分に少なければ、母平均を推論する優れた方法
  - 信頼区間を計算できる
    - 平均値を用いる場合は、ざっくり 200 事例以上
- LASSO などの機械学習は、複雑な性質を持っており、信頼区間の計算方法が確立されていない