# 統計的性質
## 機械学習

川田恵介
東京大学
keisukekawata@iss.u-tokyo.ac.jp

2025-10-07

# 1 データ分析

## 1.1 ここまでのデータ活用

- **事例ごとの個別予測** を行う予測モデルを推定する
- LASSO を活用すれば、複雑なモデルも推定できる
  - [**板橋区**、55 平米、駅から 5 分、 3 部屋、築 20 年]の物件の予測価格は 5000 万円
  - [**練馬区**、55 平米、駅から 5 分、 3 部屋、築 20 年]の物件の予測価格は 5500 万円
- 注意点: 人間行動や社会的な帰結を予測することは難しい
  - AI サギに注意 (Narayanan and Kapoor, 2024)

## 1.2 他のデータ活用

- 事例**全体**の”重要な特徴”を要約し、報告する
  - スカイラークの中期経営計画
  - 経済財政白書
- 多くの人に伝えやすいので、利害関係者が多い集団的意思決定(政策、大企業の戦略決定等)向いている
  - 経済学における実証分析の中核的な目標

## 1.3 ロールプレイ: 不動産市場分析チーム

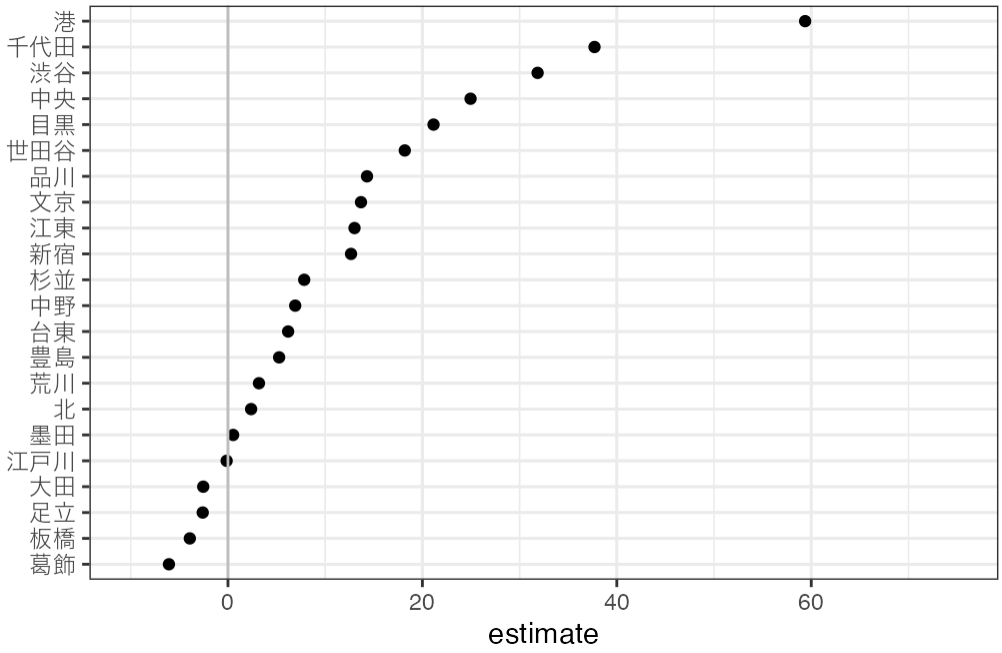
- 会社全体の意思決定を支援する情報(ファクト)として、地区ごとの平均価格を示す
- 例えば、 **データ上の** 練馬区との平均的な取引価格の差は

```
model <- lm(Price ~ fct_relevel(District,"練馬区"), data)

model
```

```
Call:
lm(formula = Price ~ fct_relevel(District, "練馬区"), data = data)

Coefficients:
                              (Intercept)
                                  33.8313
fct_relevel(District, "練馬区")世田谷区
                                  18.1896
  fct_relevel(District, "練馬区")中央区
                                  24.9547
  fct_relevel(District, "練馬区")中野区
                                   6.9161
    fct_relevel(District, "練馬区")北区
                                   2.3801
fct_relevel(District, "練馬区")千代田区
                                  37.7050
  fct_relevel(District, "練馬区")台東区
                                   6.1889
  fct_relevel(District, "練馬区")品川区
                                  14.3026
  fct_relevel(District, "練馬区")大田区
                                  -2.5434
  fct_relevel(District, "練馬区")文京区
                                  13.6837
  fct_relevel(District, "練馬区")新宿区
                                  12.6595
  fct_relevel(District, "練馬区")杉並区
                                   7.8427
  fct_relevel(District, "練馬区")板橋区
                                  -3.9237
fct_relevel(District, "練馬区")江戸川区
                                  -0.1365
  fct_relevel(District, "練馬区")江東区
                                  13.0175
  fct_relevel(District, "練馬区")渋谷区
                                  31.8522
    fct_relevel(District, "練馬区")港区
                                  59.3712
  fct_relevel(District, "練馬区")目黒区
                                  21.1468
  fct_relevel(District, "練馬区")荒川区
                                   3.1880
  fct_relevel(District, "練馬区")葛飾区
                                  -6.0715
  fct_relevel(District, "練馬区")豊島区
                                   5.2634
  fct_relevel(District, "練馬区")足立区
```

```
                                -2.5963
fct_relevel(District, "練馬区")墨田区
                                 0.5315
```

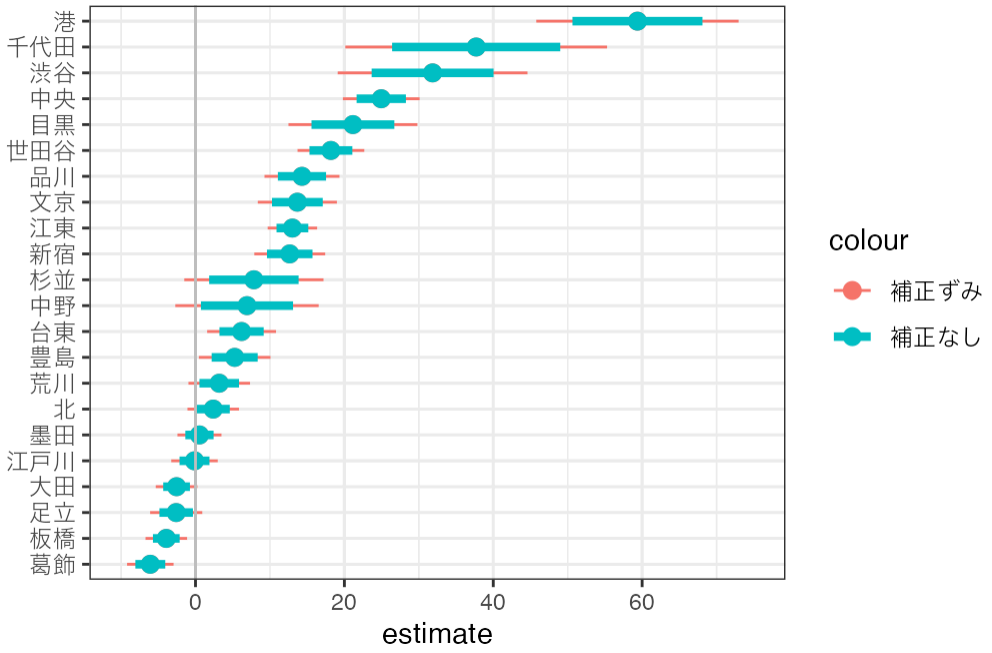
## 1.4 図示




## 1.5 本スライドの Takeaway

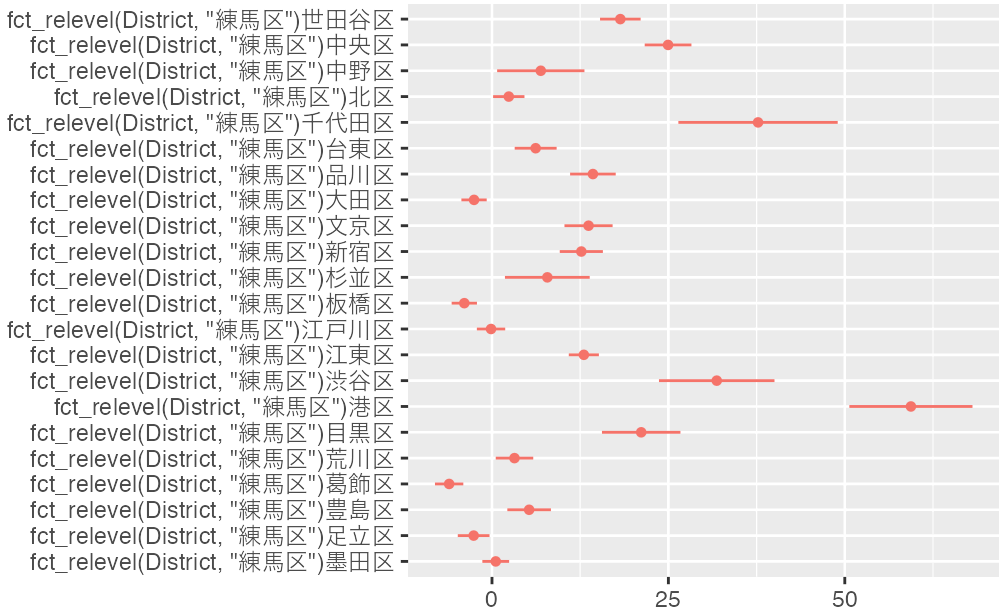
- 全ての物件(母集団)ではなく、調査された物件についての差に過ぎない
  - 母集団における平均差はどのようなものか?
- 信頼区間を用いて推論

## 1.6 信頼区間




## 1.7 推奨するコード

```
model <-estimatr::lm_robust(Price ~ fct_relevel(District,"練馬区"), data)

dotwhisker::dwplot(model)
```




# 2 コイントスの例

## 2.1 コイントス

- よく使われる例を使って、データ分析のコツを掴む
  - 研究目標: 「コイントスの結果が公平かどうかを検証しよう」
- コイントスは”公平なくじ”として、幅広く利用されている
  - 先行/後攻、昼食のお店
  - 実際にも重要な研究目標?

## 2.2 Bartoš et al. (2025)

- 研究目標 = 「投げた時と同じ面が出やすい」仮説を検証
- 350,757 回のコイントスにより、事例をサンプリング
  - 実際のサンプリング風景の Youtube
  - 母集団 = 無限回のコイントスを行った結果

## 2.3 Bartoš et al. (2025)

- データ上で、同じ面になった回数は、178,079 回
  - $178079/350757 \approx 0.507$
    - = 推定値

- 同じ面が出やすい傾向

- 批判: 偶然であり、もう一回実験をやれば結果が変わるのでは?

## 2.4 イメージ

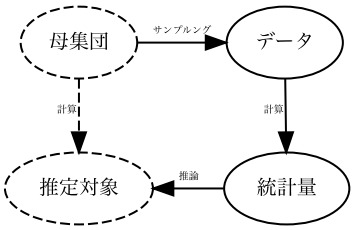



Figure 1: Elephant

## 2.5 再現可能性

- コツは、「得られた可能性があるが、実際には得られていないデータ」を想像すること
- 二人が、5 回コイントスを行いデータ収集
  - 全員のデータが異なるので、結果も異なる
    - 結果が再現できないので、誰の結果も信頼できない
- あくまでフィクションであることに注意
  - 本当に各人がデータを収集したのであれば、データを共有し、結合すべき

## 2.6 例

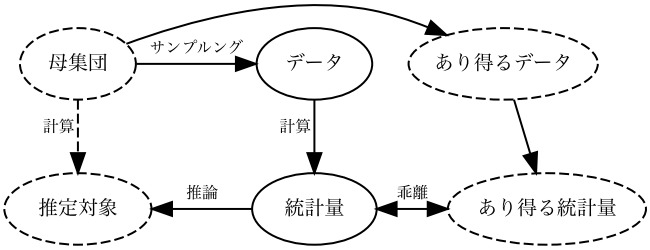



Figure 2: Elephant

## 2.7 あり得た結果

- 自身の研究目標は、基本的に「自分以外は行わない」ケースが多い
  - **仮想的な** 別のデータを常に意識する必要がある
    - 頻度論と呼ばれる枠組み

## 2.8 統計的推論

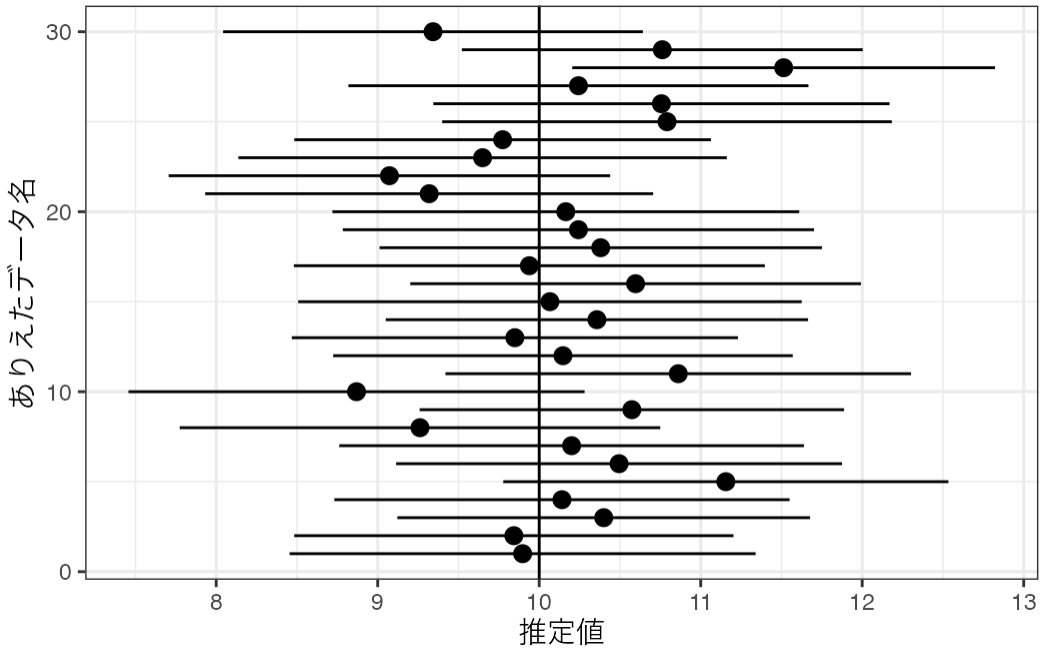
- 同じ研究目標に対して、さまざまなデータを得る可能性があるので、確実に正しい結論を得ることは、**あきらめる**

- 概ね正しい結論を目指す
  - 主張 「母集団において同じ面になる割合は、**概ね(95% の確率で)** 信頼区間 $[0.506, 0.509]$ の間」は真
  - 事例がランダムサンプリングされており、十分な事例数 (150 事例以上)が存在することが前提

## 2.9 信頼区間

- データから計算される区間
  - 同じ母集団、推定対象であったとしても、データが異なれば、異なる区間が計算される
    - ただし多くのデータ (初期値では、潜在的に生成されうるデータの 95 %) について、推定対象を含む区間が算出される
    -「自身が計算した信頼区間は、概ね、推定対象を含む」

## 2.10 データ




## 2.11 非実用的な推論の例

- 前提
  - 無限回のコイントス(無限大の事例)から得られた同じ面になる割合は、0.508
- 結論 = 母集団において、同じ面になる割合は、0.508
  - 一致性と呼ばれる性質

- 問題点: 無限大の事例を持つデータは、存在しない

## 2.12 非実用的な推論の例

- 前提
  - 100 % 当たる予測モデルが使える
  - 同じ面が出る予測確率は、0.508
- 結論 = 母集団において、同じ面になる割合は、0.508
- 問題点: 確実に当たるモデルは、(社会/市場分析では)想定しずらい

## 2.13 Reference

# Bibliography

Bartoš, F. et al. (2025) “Fair coins tend to land on the same side they started: Evidence from 350,757 flips,” Journal of the American Statistical Association, pp. 1–10.

Narayanan, A. and Kapoor, S. (2024) “AI snake oil: What artificial intelligence can do, what it can’t, and how to tell the difference,” AI Snake Oil. Princeton University Press.