誤差

労働経済学 2

川田恵介

データ分析の根本問題

- データ $\neq$ 関心のある社会
  - 関心のある社会を調べ尽くすことは” 不可能”
  - 調査者によって結論が異なる
- 例: クラス全員が、**それぞれ独立して** 、池袋駅利用者調査を行う
  - 利用者男女比率についての調査結果は？

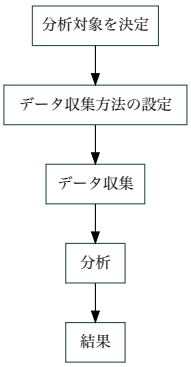
注意が必要な文言

- “統計学的には 2 千人の回答があれば、「**有権者全体の縮図**」として**十分な分析**ができる程度の**誤差**におさまるといわれてます”
  - 朝日新聞の世論調査より

再現可能性

- ある主張の**正しさ**をどのように検証？？？
  - 自分の目で確かめる
  - 少なくとも独立した第 3 者によって検証できる
- 一部自然科学では可能
  - 純粋な水は誰がやっても、１００度で沸騰する
- 個体差が大きな事象を対象にする分野 “Soft Science” (医学、経済学、政治学などなど) では困難

データ分析ロードマップ



ランダム・サンプリング

- 最も**理想的**なデータ収集方法
  - 調査対象を知りたい集団 (母集団) から、ランダムに選ぶ
  - 多くの公的統計で活用
- **正しい** 手法を使えば、知りたい集団の特徴を正しく推論できる
  - 厳密な再現可能性は引き続きない

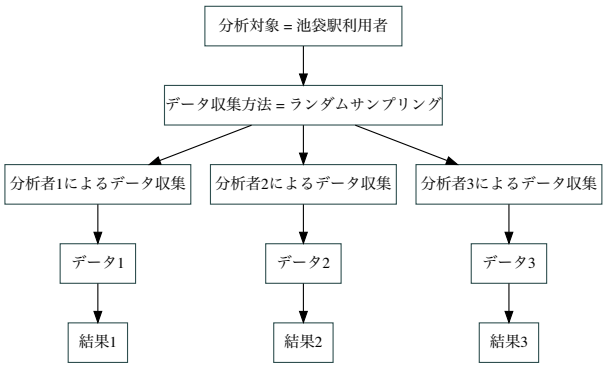
何を問題とするのか

- 分析対象が異なれば、結果が異なるのは当たり前
- データ収集方法が異なれば、結果が異なりうる
  - 早朝の池袋で調査 VS 深夜の池袋で調査
- 分析手法が異なれば、結果が異なる

  - 分析の頑強性

- **データ収集方法、分析対象、手法が同じでも結果が異なる**

データ分析ロードマップ



例

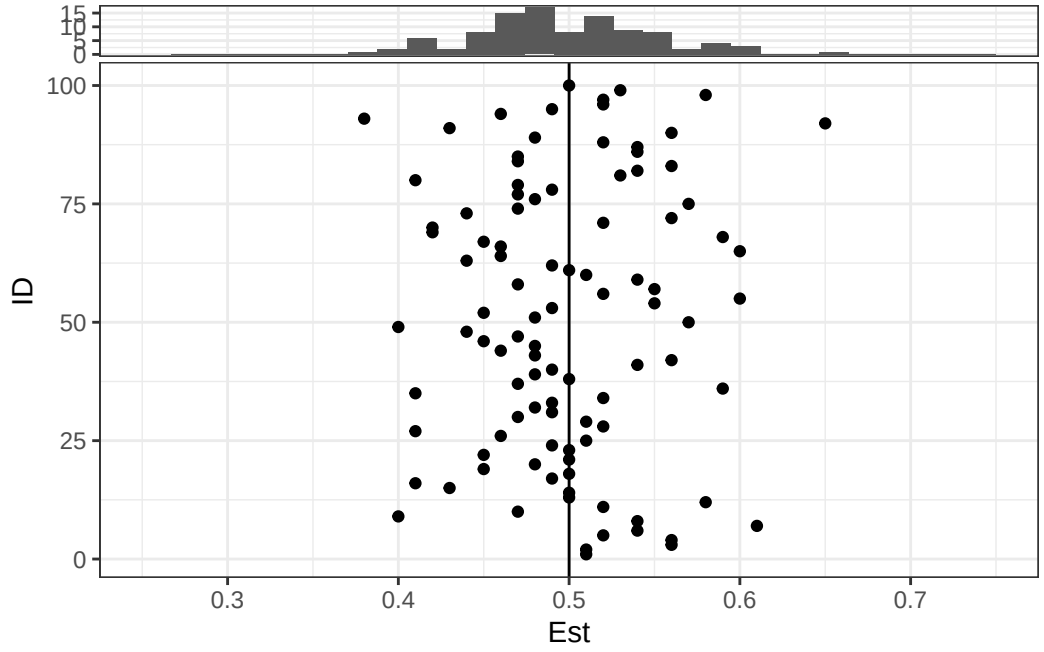
- 各分析者が 100 人または 10 人を調査
  - 本当の男女比率は 50%
- ランダムサンプリングを行う

数値例: 100 サンプル

```
Registered S3 method overwritten by 'ggside':
  method from
  +.gg   ggplot2
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

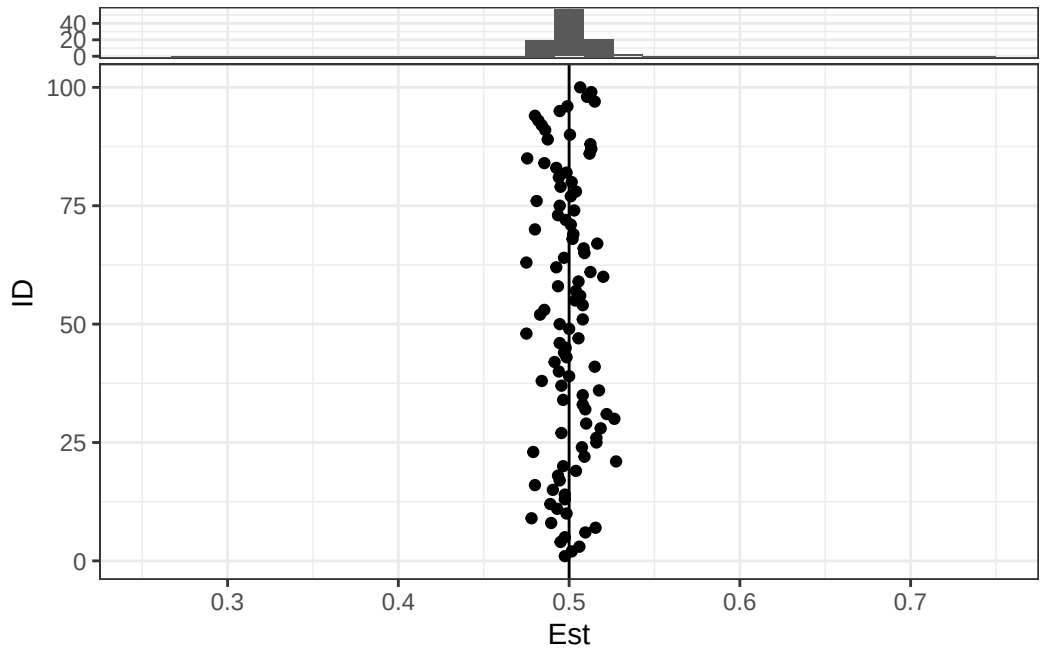
```
Warning: Removed 1 rows containing missing values (geom_xsidebar).
```



数値例: 2000 サンプル

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 1 rows containing missing values (geom_xsidebar).
```



ランダムサンプルデータの性質 1

- サンプルサイズが無限大に大きくなると、真の値 (母集団における平均) と一致
  - 経済学研究ではほぼ実用性がない
  - かなり大きなデータでも、“致命的” な誤差が生じうる
  - 失業率が 5% ずれたら、、、、？
- 対策: 完全一致は諦める！！！

信頼区間

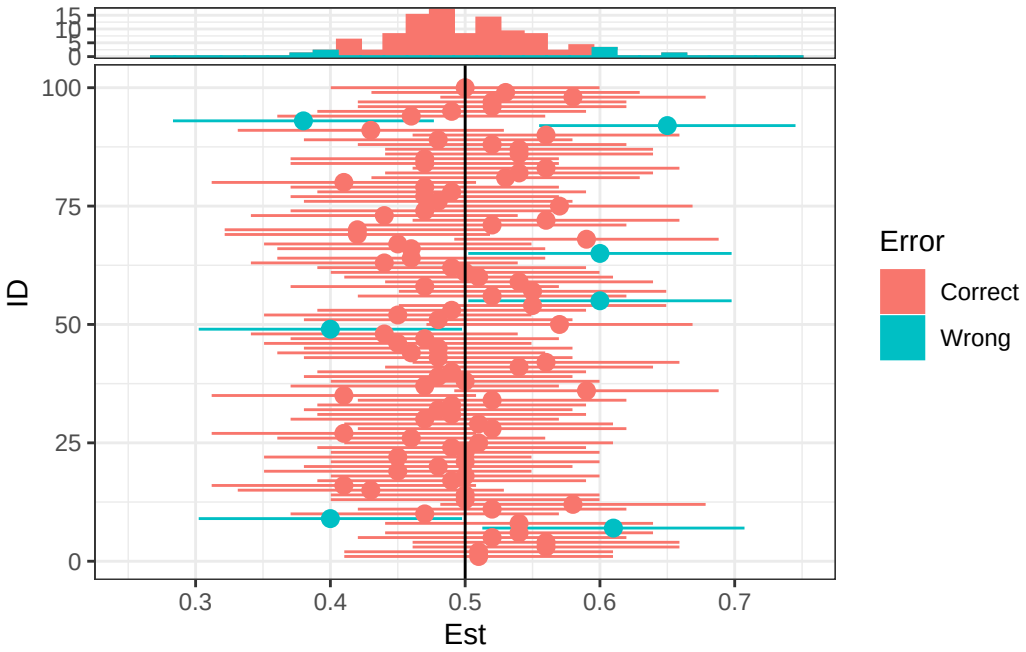
- 区間で答える
- 研究者の内 $\alpha\%$ が、真の値を含んだ区間を得られる

数値例: 100 サンプル

‘

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 2 rows containing missing values (geom_xsidebar).
```

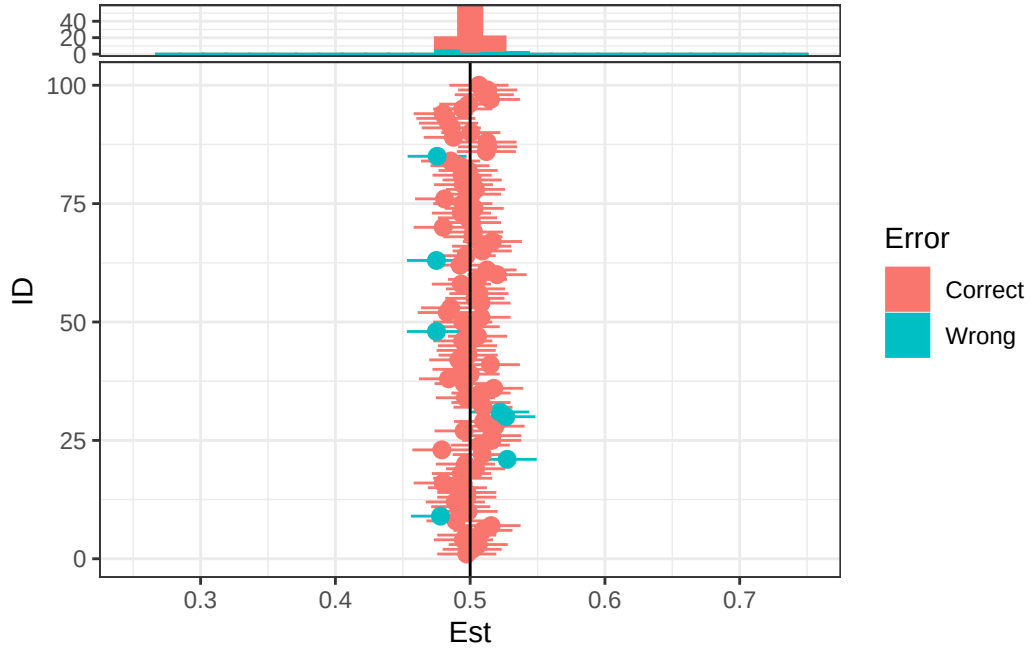


数値例: 2000 サンプル

‘

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 2 rows containing missing values (geom_xsidebar).
```



算出方法

- 複数の方法が存在
- 代表的な方法は、漸近正規性 (サンプルサイズがある程度聞くなると、推定値は正規分布に近づく)
- 個票データがあれば、容易に計算可能
  - 計量経済学や機械学習の講義を取ってください
- 割合については、個票データがなくても計算可能

計算方法

- $\hat{p}$ = データ上の割合; N = サンプルサイズ
- 95% 信頼区間

$$[\hat{p} - 1.96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{N-1}}, \hat{p} + 1.96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{N-1}}]$$

他の指標との関係

- p 値,t 値: 母平均が特定の値であるかどうかを検定
- 信頼区間と同じ理屈から算出
- “0” が 95% 信頼区間の外であれば、p 値は 0.05 以下

まとめ

- データと現実には常にずれがある
- 全く同じ手続きを下手としても、ランダムサンプリングである限り、分析結果は一致しない
- 点ではなく、線で論じる