推定

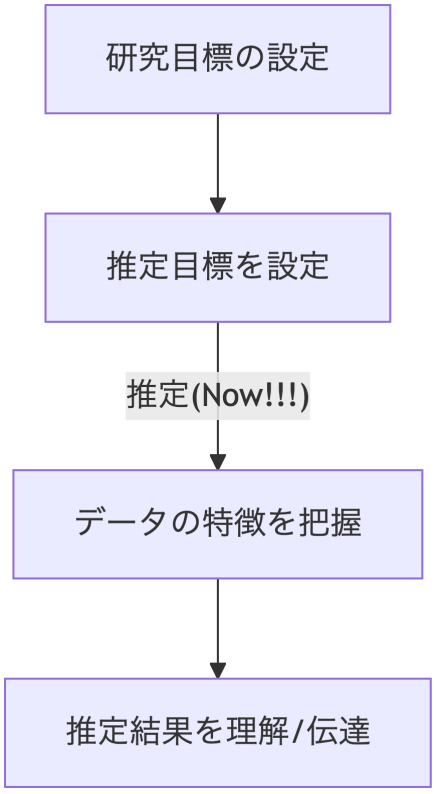
Data visualization

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

2025-07-30

1 推定

1.1 Work flow



1.2 根本的な疑問

- データから、社会/市場について、何がわかる？

- 何がわからない?

1.3 再現可能性

- 重要な”自問”: 同じ手法でデータ収集/分析を行った時に、同じ結果を得ることができるか?
  - データ取集: インタビュー、郵送、インターネット、電話調査
- 事例の選ばれ方に”ランダム性”があり、データが異なる
  - 同じ結論を得ることができない
- 同じ結果を得られないのであれば、目の前の結果を信じる合理的理由が存在しない
  - データ $\neq$ 社会の姿

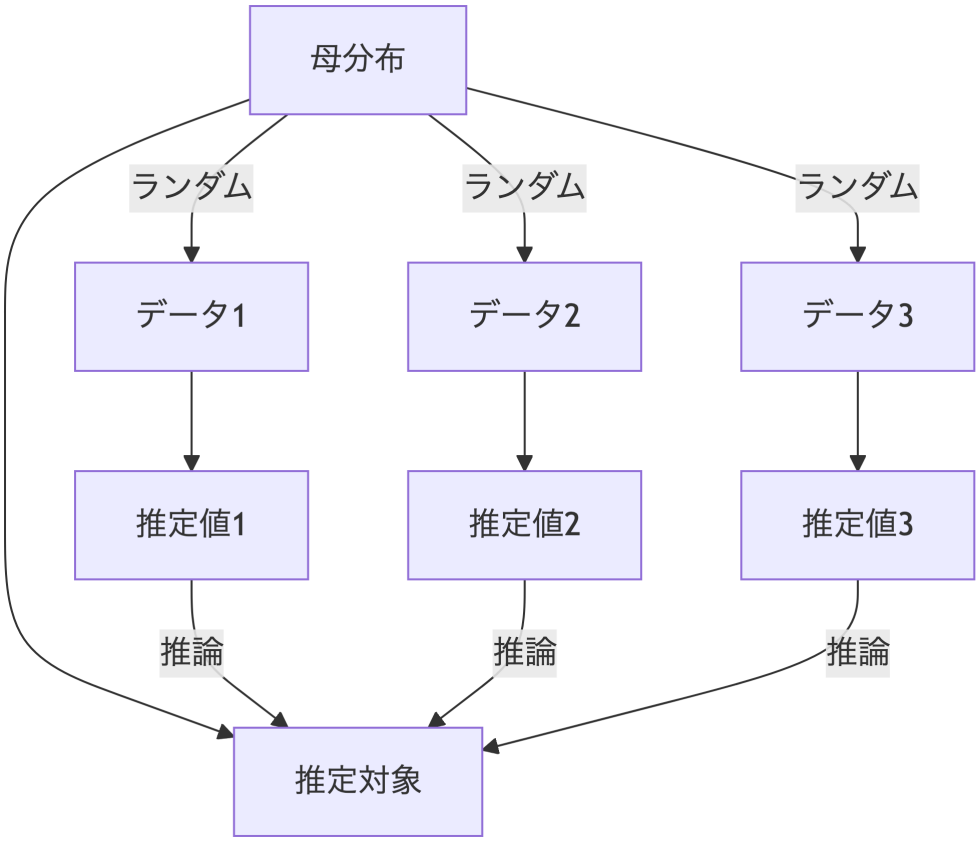
1.4 Key concepts: 母集団

- 再現可能性を議論するために、母集団を導入
  - 本講義の範囲内では、**研究対象である巨大なデータ**をイメージして OK
- データの事例は、研究関心となる母集団からランダムサンプリングされていると仮定

1.5 推定対象

- 母集団上で、**仮想的に計算される**、分布や平均値
  - 「正答(母集団の特徴)は共通だが、データから得られる回答は異なる」状況をイメージ

1.6 イメージ

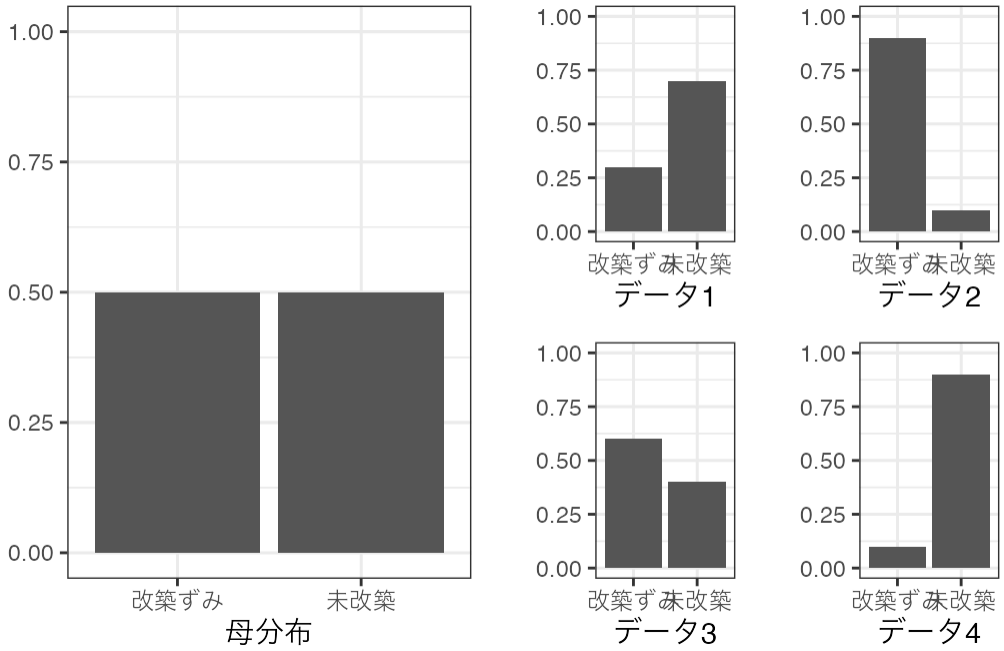


2 母分布の推定

2.1 母分布

- データと同様に、母集団も、母分布に集約できる
  - 単純なケースであれば、推定対象にできる
- 例: 推定目標 = 改装された物件の割合
  - 研究目標 = 東京全体で改装された物件の割合を知りたい

2.2 母分布とデータの分布



2.3 推定のアイディア

- データ数が十分にあれば、データ上での割合 ≃ 母集団上での割合、が期待できる
  - 一致性
- データ上の改装ずみの割合 = [$Y = 1$ であれば改装ずみ、$Y = 0$ であれば未改装]、の平均値

```
mean(data$Reform)
```

```
[1] 0.2540005
```

2.4 サンプリング誤差

- データ上での割合 = 母集団上での割合 ではない
  - 現実的な事例数では、データの性質と母集団の性質は必ず乖離する
    - 改装ずみの物件を偶然過剰/過小に収集する可能性がある
- サンプリング: データの事例を収取する方法
  - 代表例: ランダムサンプリング
    - 事例を母集団からランダムに収集する

2.5 数値実験

- 改装ずみの物件は、母集団において 5 割
  - ランダムに 5 事例を抽出

```
Y <- sample(0:1, 5, replace = TRUE)
Y
```

```
[1] 0 1 1 0 0
```

2.6 数値実験

- 平均値を計算すると

```
mean(Y)
```

```
[1] 0.4
```

- 再度データを収集し直すと

```
Y <- sample(0:1, 5, replace = TRUE)
mean(Y)
```

```
[1] 0.8
```

2.7 繰り返し実験

- 10 回、5 事例のデータを抽出

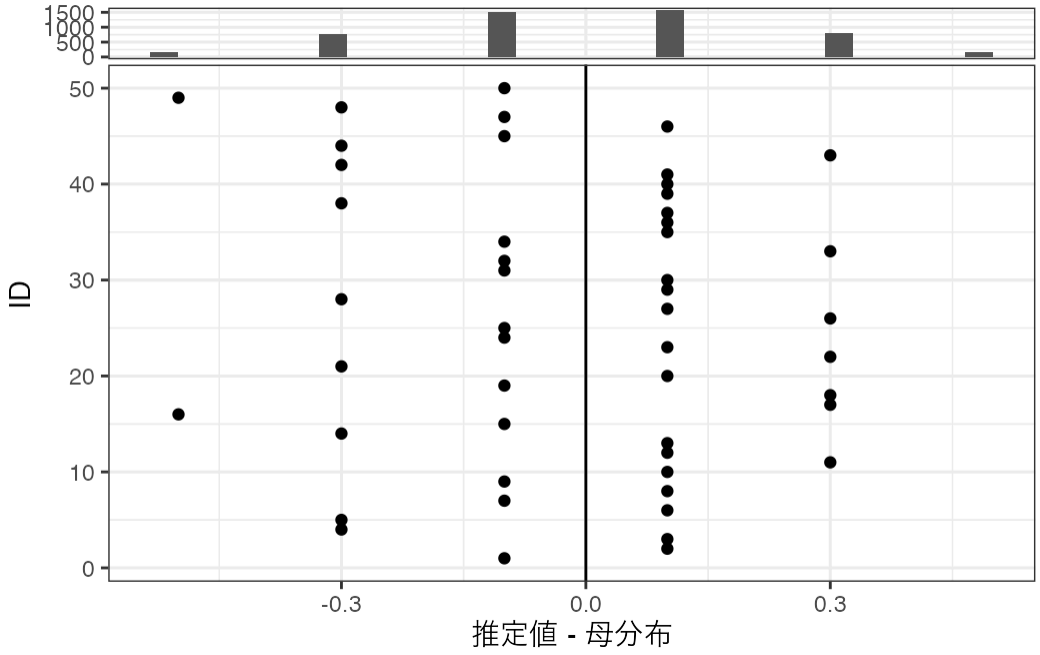
```
for (i in 1:10){
  set.seed(i)
  Y <- sample(0:1, 5, replace = TRUE)
  Mean <- mean(Y)
  print(Mean)
}
```

```
[1] 0.4
[1] 0.6
[1] 0.6
[1] 0.2
[1] 0.2
[1] 0.6
[1] 0.4
```

```
[1] 0.6
[1] 0.4
[1] 0.6
```

- 推定値の分布と呼ばれる

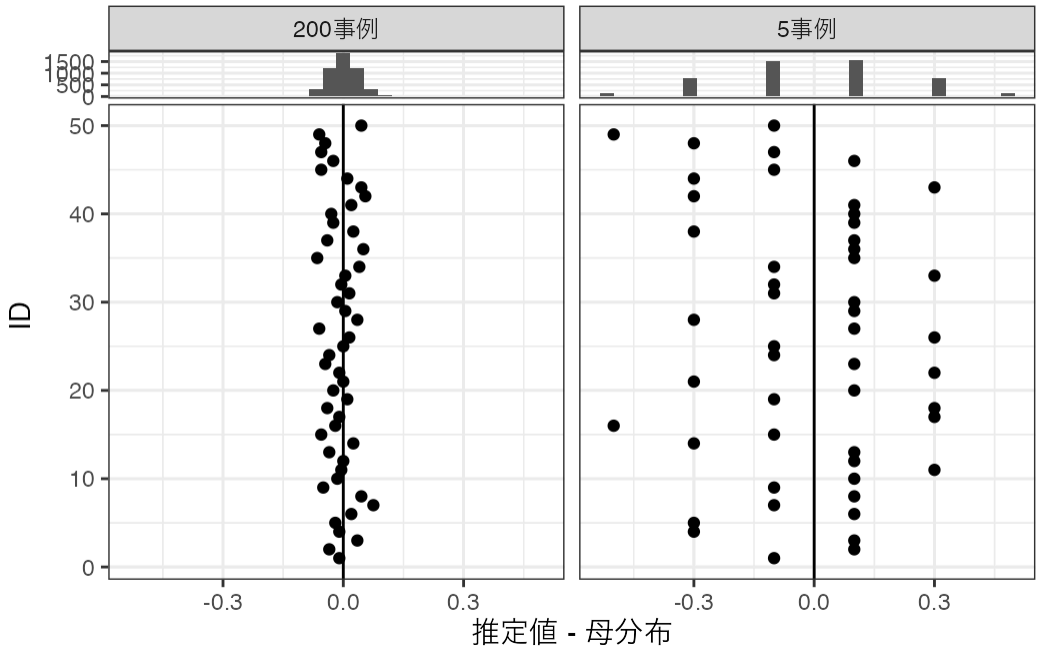
2.8 繰り返し実験



2.9 サンプリング誤差の性質

- ランダムサンプリングデータであれば、事例数が増えるにつれて、母分布に近づく
  - 一致性

2.10 繰り返し実験: 200 事例



2.11 サンプリング誤差の性質

- データの分布が母分布と正確に一致するには、“無限大”の事例数が必要
  - 「データの分布 = 母分布」は、100 % 誤った主張

2.12 統計的推論

- 仮定 (本スライドでは、データがランダムサンプリングされている)のもとで、**ほぼ**正しい主張を導きたい
- 代表的なものは、信頼区間
  - 一定確率 (初期値では 95%) で、母分布を含む区間

2.13 統計的推論: 例

```
library(tidyverse)

data <- read_csv("Data/example.csv")

estimatr::lm_robust(Reform ~ 1, data)
```

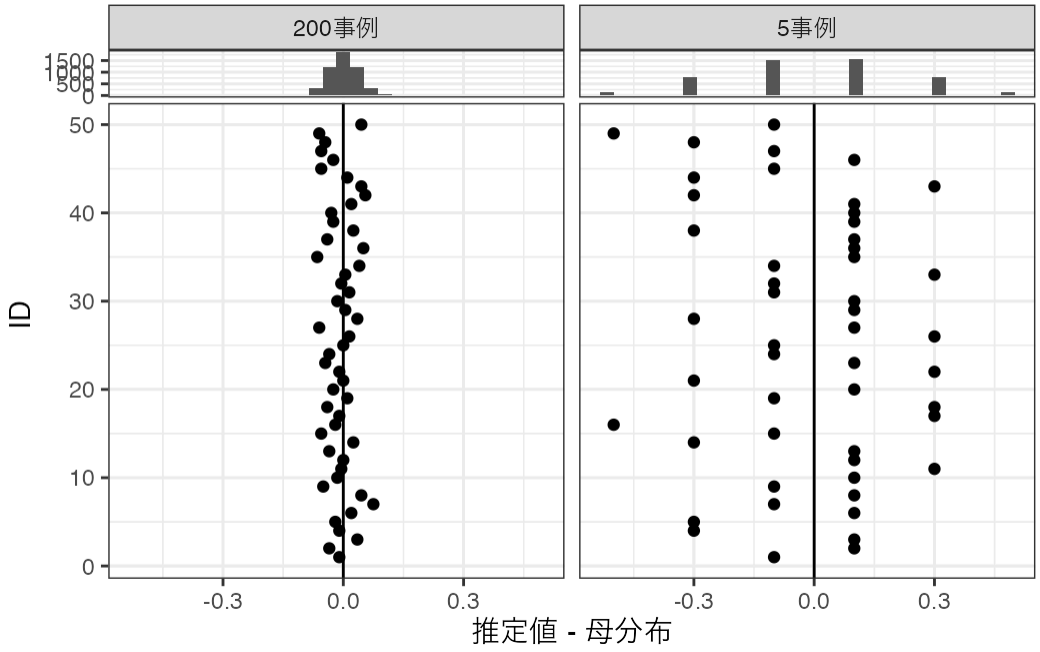
```
             Estimate  Std. Error  t value Pr(>|t|)  CI Lower  CI Upper    DF
(Intercept) 0.2540005 0.004093127 62.05537        0 0.2459773 0.2620238 11310
```

- $[0.246, 0.262]$ は、概ね母集団における改装ずみの物件割合を含む

2.14 信頼区間の根拠

- ある程度の事例数があれば、分布の推定値 (データ上の平均値)は、正規(ベル/富士山型)分布で近似できる性質を利用
  - 中心極限定理
- 150 事例程度以上あり、値が極端に偏っていないことを要求

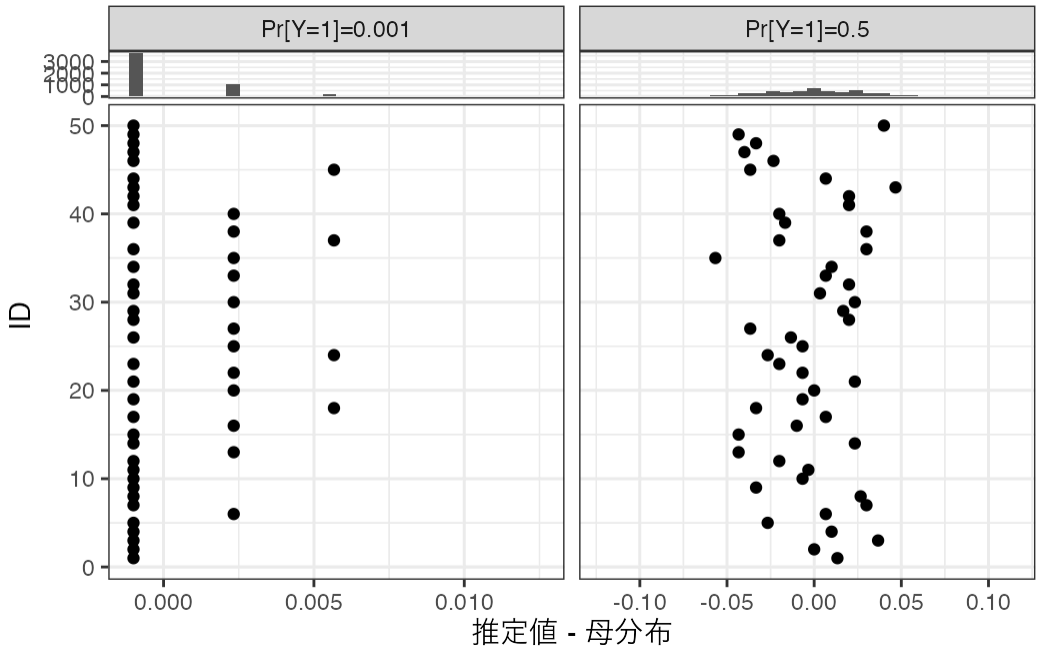
2.15 繰り返し実験: 200 事例



2.16 分布の推定の問題

- 複数の変数についての母分布の推定は難しい
- 例: $[Size, Tenure]$ について、母分布を推定
  - $Size = 50, Tenure = 10$ であれば、$Y = 1$、それ以外であれば $Y = 0$ となる変数を定義すると、ほぼ”0”となる
  - 母分布との乖離が激しく、中心極限定理が適用できない

2.17 数値例



2.18 まとめ

- 母分布の直接的な推定は、特殊なケース (単純な母分布) でのみ可能
  - 多くの応用では、異なるアプローチを用いる

3 集計値の推定

3.1 母平均

- 母分布の推定が難しい場合、分布のシンプルな特徴を推定目標とすることが有益
- 代表例は、母平均
  - Price の分布ではなく、**母集団上での仮想的に計算した** Price の平均値

3.2 母平均の推定

- アイディアは同じ
  - データ上の平均値 $\simeq$ 母平均
  - 信頼区間の計算も可能

3.3 例

```
library(tidyverse)
```

```r
data <- read_csv("Data/example.csv")

estimatr::lm_robust(Price ~ 1, data)
```

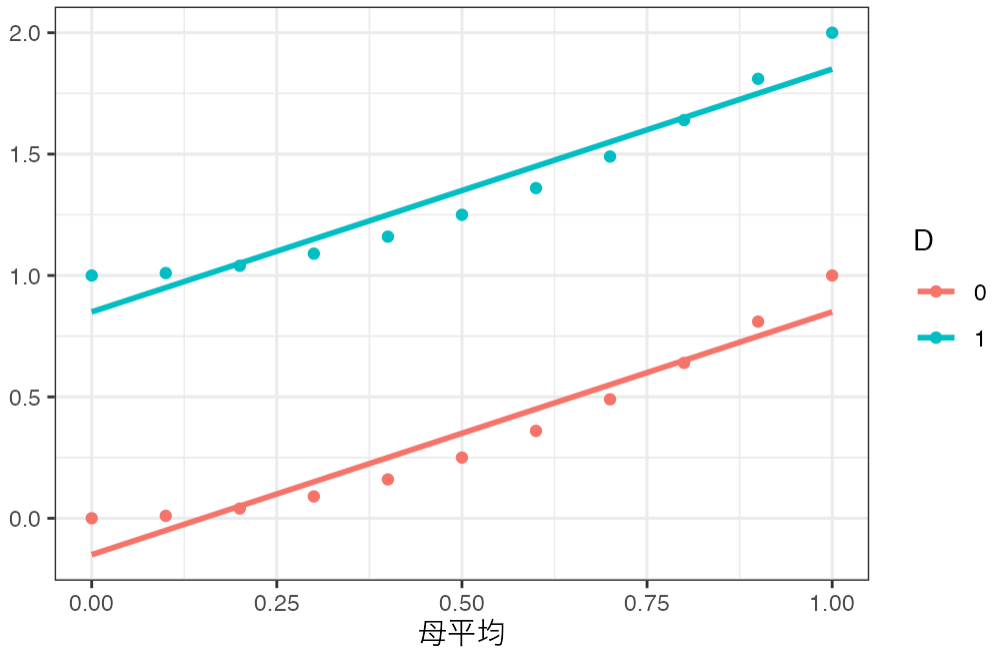
```
            Estimate Std. Error  t value Pr(>|t|) CI Lower CI Upper    DF
(Intercept) 45.24285  0.4252469 106.3919        0 44.40929 46.07641 11310
```

3.4 Population OLS

- **母集団上での仮想的に計算した** OLS の結果も推定できる
  - 母集団上での Price と Size の関係性を把握するために、
    - 母集団上で Price を Size で回帰した結果、を推定する

3.5 例. Populaiton OLS

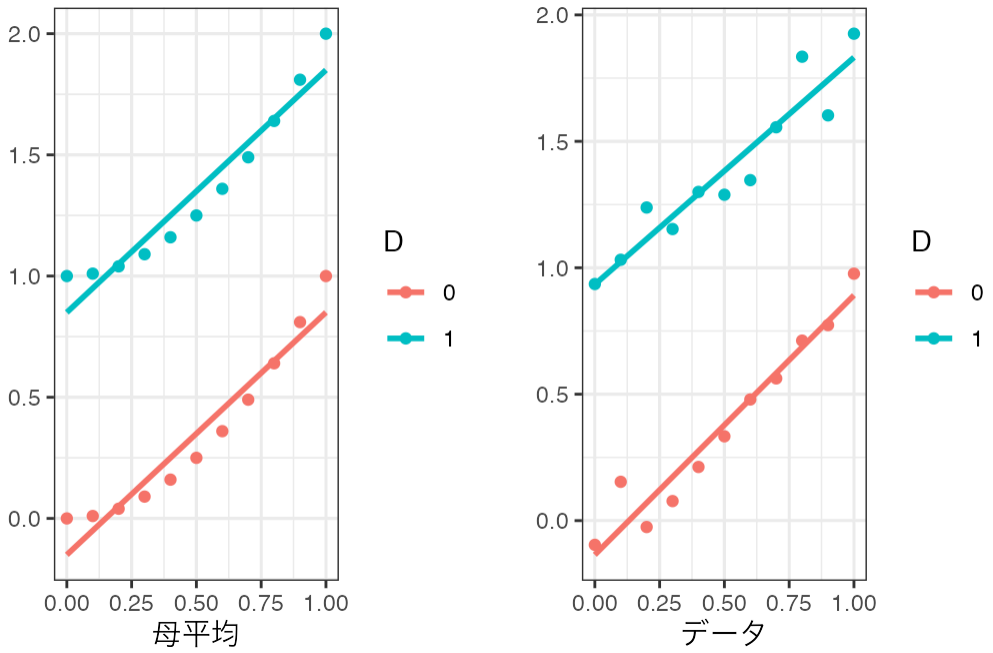
- $Y \sim D + X$



3.6 Population OLS の推定

- アイディアは同じ:
  - データ上での OLS ≃ 母集団上での OLS
  - 信頼区間の計算も可能

3.7 例. OLS



3.8 例

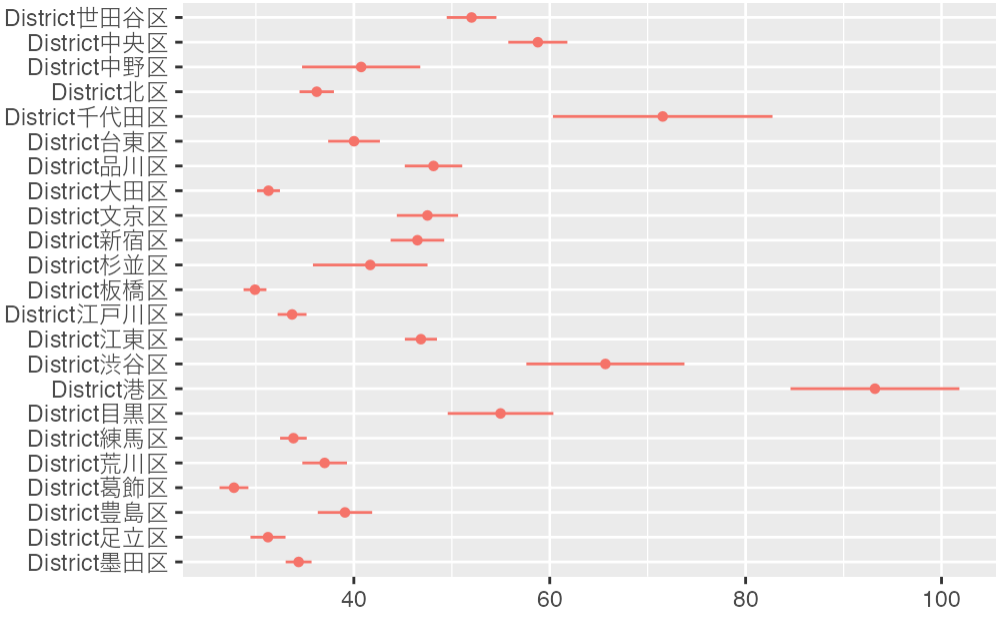
```
library(tidyverse)

data <- read_csv("Data/example.csv")

estimatr::lm_robust(Price ~ Size + year_2024, data, alpha = 0.05)
```

```
              Estimate Std. Error   t value      Pr(>|t|)   CI Lower   CI
Upper
(Intercept) -13.372975 1.45248032 -9.206992  3.942895e-20 -16.220089
-10.525861
Size          1.130698 0.03321124 34.045656 5.954754e-242   1.065598
1.195798
year_2024    14.478989 0.69799824 20.743589  7.733393e-94  13.110791
15.847187
               DF
(Intercept) 11308
Size        11308
year_2024   11308
```

3.9 例

```
model <- estimatr::lm_robust(Price ~ 0 + District, data)
```

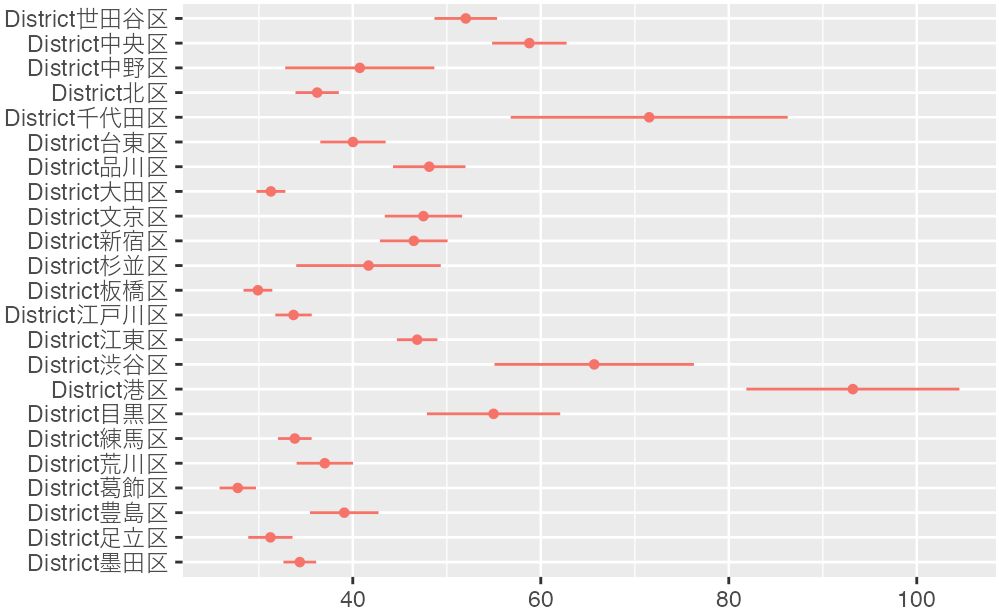
```
dotwhisker::dwplot(model, ci = 0.95)
```



3.10 信頼水準の変更

- 信頼水準 = 信頼区間が母集団上で計算した値を含む確率
  - 0.95 から変更できる

```
model <- estimatr::lm_robust(Price ~ 0 + District, data, alpha = 0.01)

dotwhisker::dwplot(model, ci = 0.99)
```



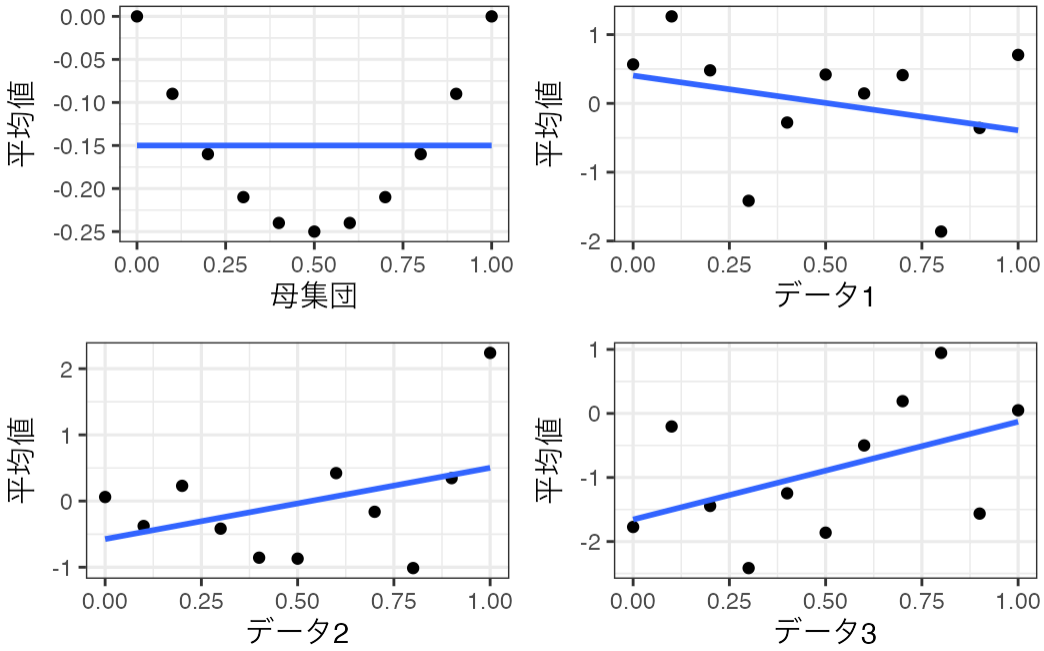
3.11 絶対に正しい結論

- 信頼水準を 100 % に設定できるか?
  - 信頼区間が無限大に広くなる
    - 結論が、「平均取引価格は$-\infty, \infty$ の範囲内」に変化
      - **何も述べないので**、絶対に正しい!!!
        - ???

3.12 誤定式化の下の推定

- OLS は、あくまでも、Population OLS の結果を推定している
  - 誤定式化の下では、母平均を推定しているわけではない

3.13 例. OLS



3.14 Takeaway

- データではなく、データの背後にある母集団を推定したい
  - 母分布の推定は難しいので、母分布の特徴(OLS や平均値)の推定を目指すことが現実的
  - 信頼区間によるサンプリング誤差の影響を定量化することが重要

Bibliography