データの特徴を把握

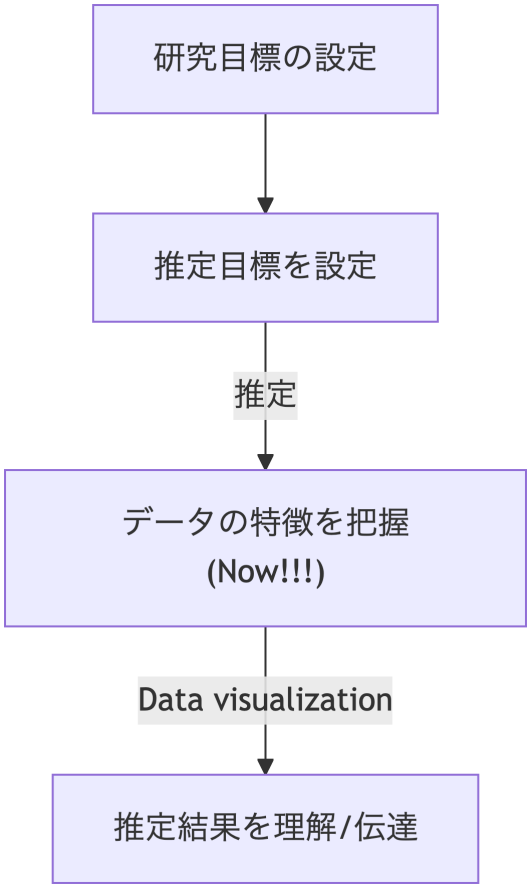
Data visualization

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

2025-07-30

1 データ

1.1 Work flow



1.2 データ例

District	Price	Size	year_2024
千代田区	94	40	1
千代田区	100	65	0
千代田区	130	65	1
千代田区	98	65	0
千代田区	58	40	0
千代田区	330	95	1

1.3 個別事例分析

- 最も取引価格が高い物件は以下

District	Price	Size	year_2024
杉並区	1400	105	1

- “2024 年に取引された杉並区の 105 平米の物件は、14 億円で取引される”と一般化可能?

1.4 個別事例分析

District	Price	Size	year_2024
杉並区	93	105	1
杉並区	1400	105	1

- データ上、全く同じ特徴を持つが、取引価格が大きく異なる事例が存在
  - 特殊な事例のみに注目すると、データの持つ情報の多くを捨ててしまい、ミスリードな印象を与えてしまう

1.5 分布

- 事例の並び順がランダムに決まっているのであれば、データの持つ情報は、各変数の組み合わせの割合に、“完全に” 集約できる
  - “分布”と呼ばれる

1.6 例. Size の分布

Size	N
15	638

Size	N
20	1913
25	1339
30	451
35	417
40	599
45	423
50	682
55	910
60	846
65	948
70	893
75	462
80	295
85	160
90	93
95	56
100	53
105	133

1.7 表の限界

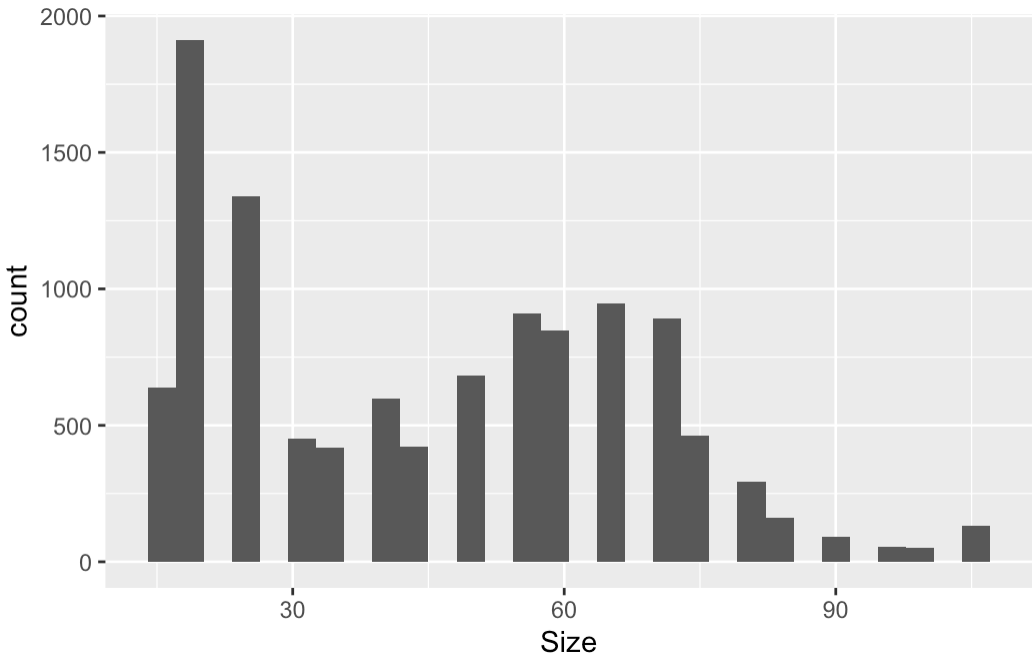
- X の数が増えたり、X の中に大量の値をとる”連続”変数 (例: 年齢, 所得)が含まれている場合、巨大な表が必要になり、人間が理解できなくなる
  - $X = [Size, District] = 416$ 行が必要
  - $X = [Size, District, Price, Distance, District] = 6565$ 行が必要

1.8 ヒストグラムによる可視化

- 代表的な方法は、ヒストグラムの活用
  - 変数を適当に区切り、対応する事例数を縦軸に表示する
- R などでは、グループごとにヒストグラムを書くことも容易にできる

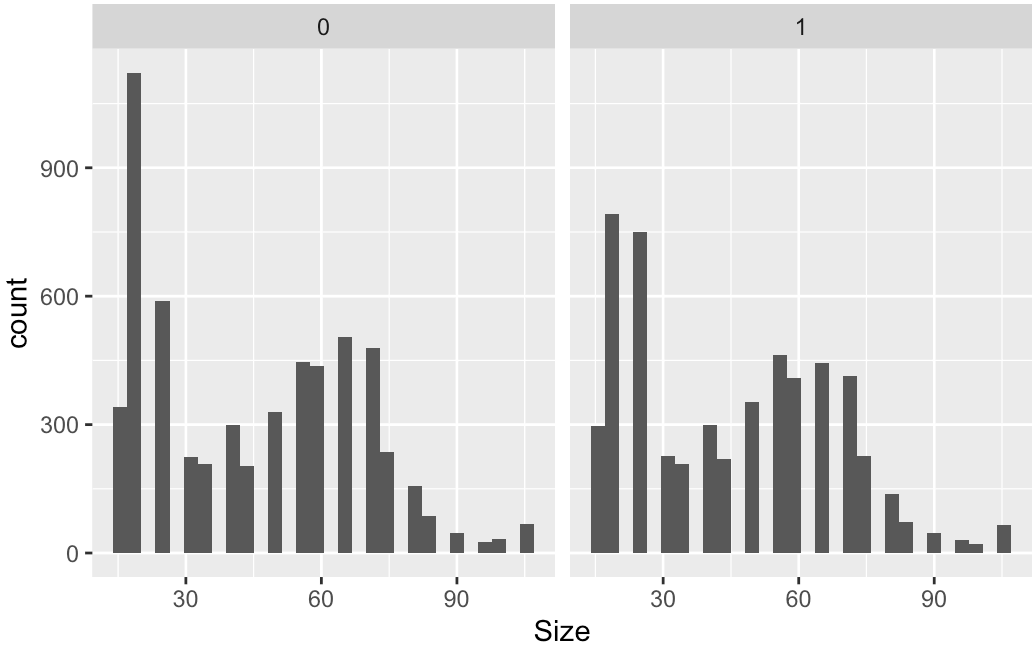
1.9 例: Size

```
data |>
  ggplot(
    aes(x = Size)
  ) +
  geom_histogram()
```



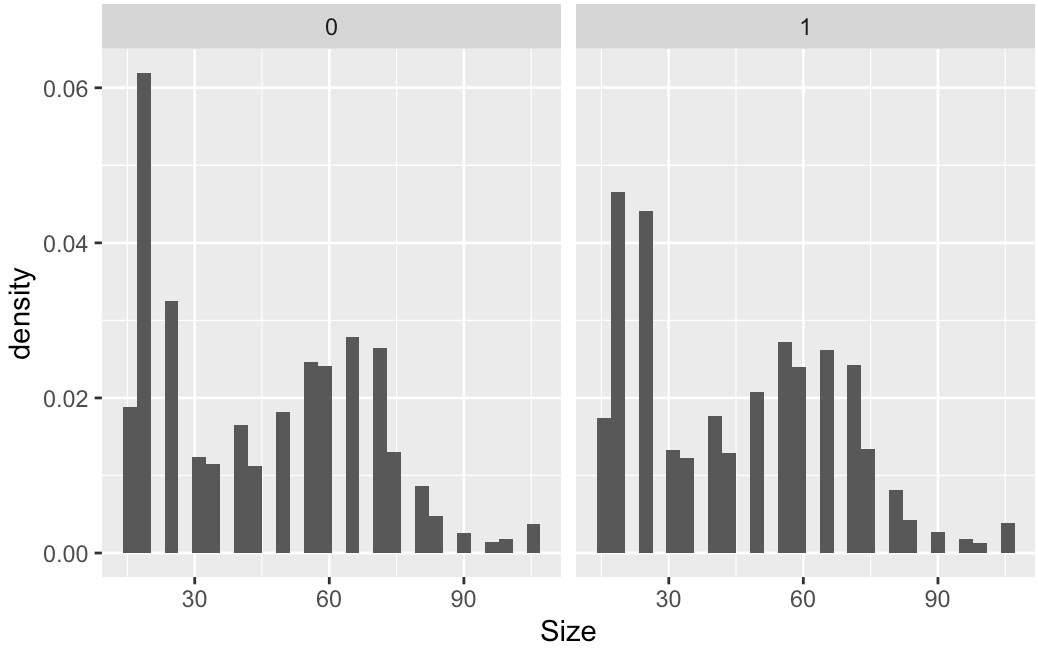
1.10 例: Size,year_2024

```
data |>
  ggplot(
    aes(x = Size)
  ) +
  geom_histogram() +
  facet_wrap(~ year_2024)
```



1.11 例: Size,year_2024

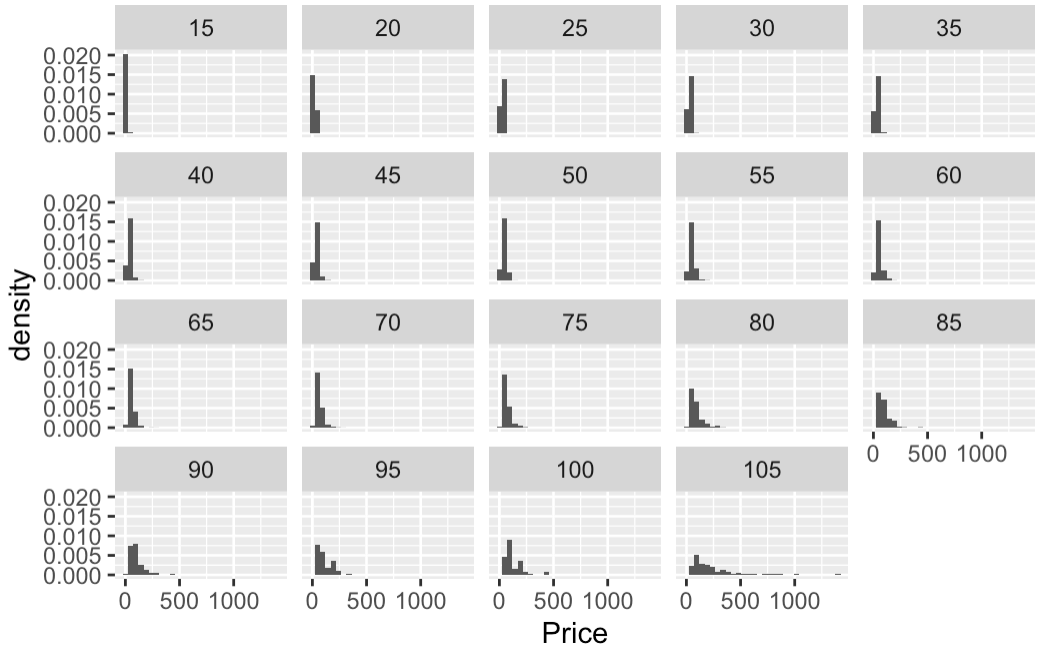
```
data |>
  ggplot(
    aes(x = Size)
  ) +
  geom_histogram(
    aes(
      y = after_stat(density)
    )
  ) +
  facet_wrap(~ year_2024)
```



1.12 ヒストグラムによる可視化の限界

- 複数の連続変数の分布を表現するのが難しい

1.13 例: Price,Size

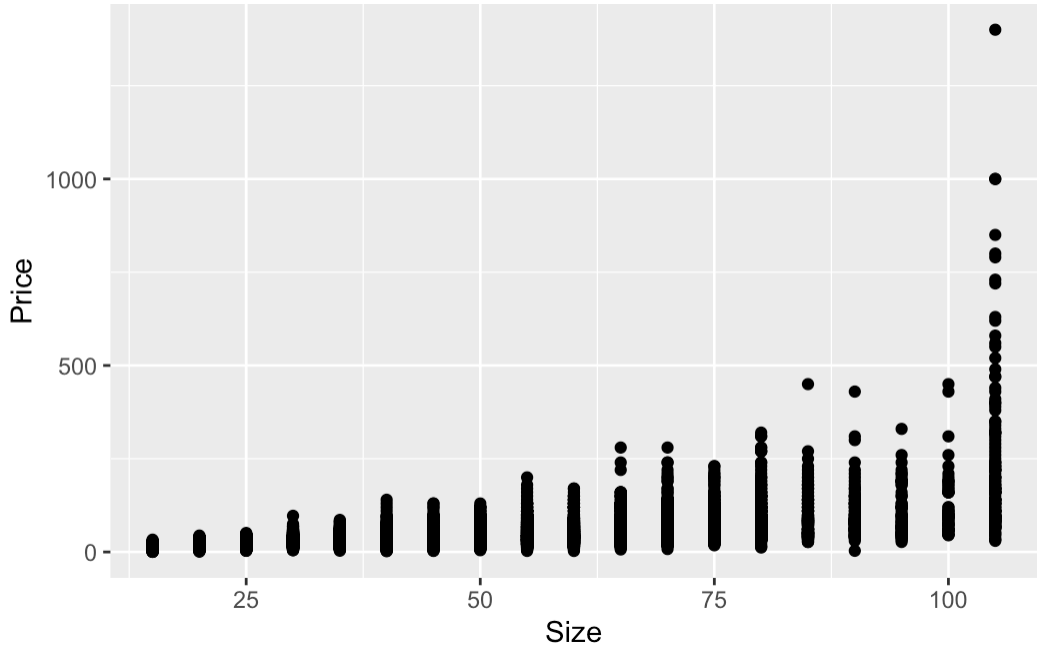


1.14 散布図による可視化

- ２つの変数の値を点で表す
- 発展: ヒートマップ
  - 事例数を色で示す

1.15 散布図: Size,Price

```
data |>
  ggplot(
    aes(
      x = Size,
      y = Price
    )
  ) +
  geom_point()
```

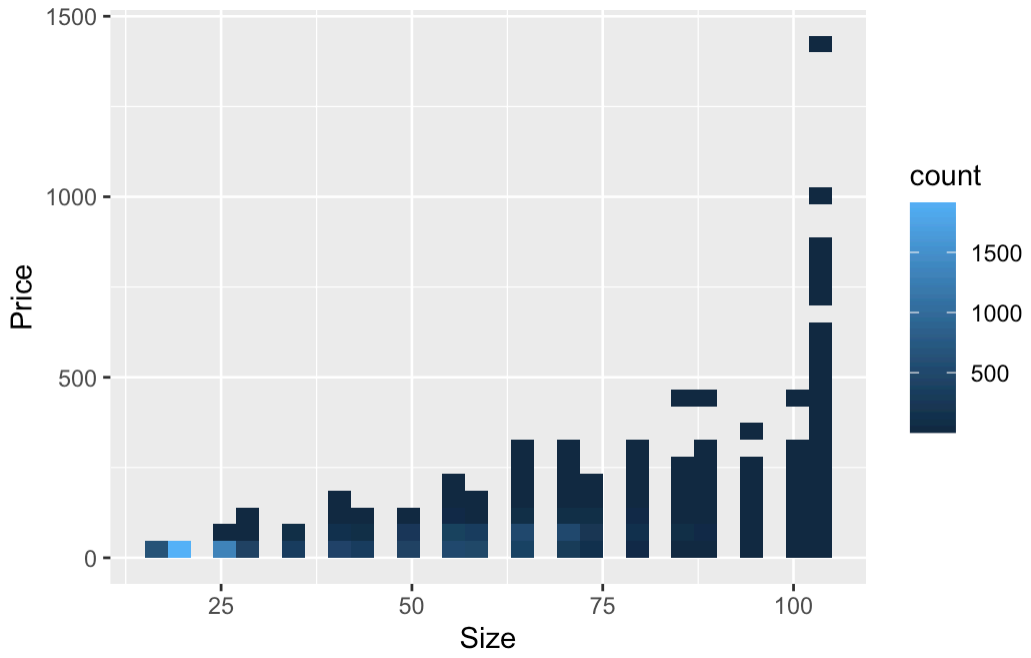


1.16 散布図の限界

- 事例数が多いデータでは、“点が潰れてしまい”、分布が認識しにくくなる
- 解決策: ヒートマップ
  - 事例数を色で示す

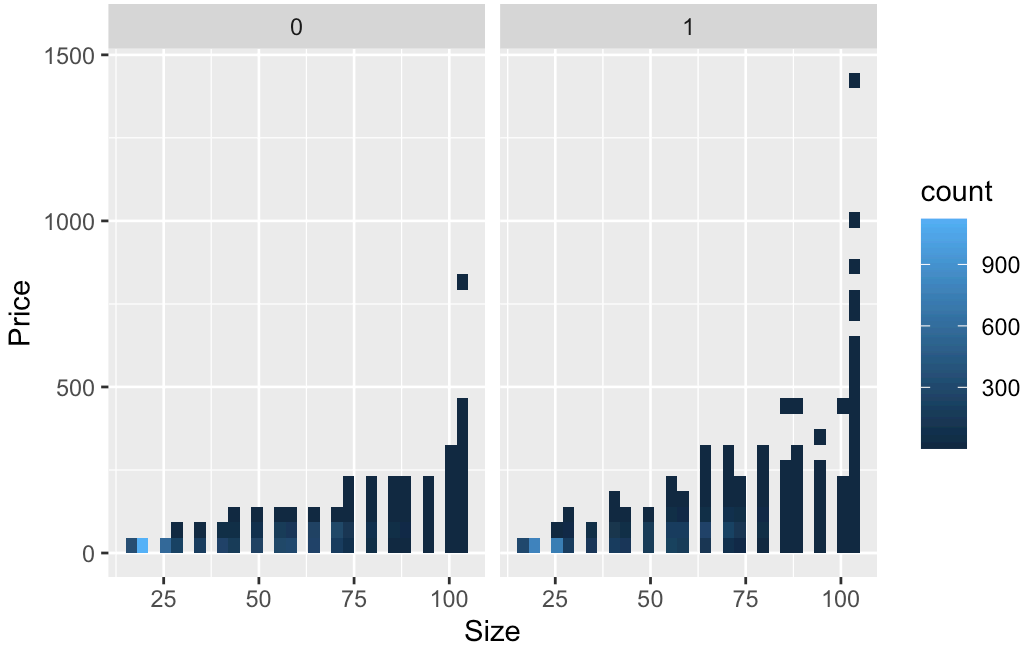
1.17 ヒートマップ: Size,Price

```
data |>
  ggplot(
    aes(
      x = Size,
      y = Price
    )
  ) +
  geom_bin2d()
```



1.18 ヒートマップ: Size,Price,year_2024

```
data |>
  ggplot(
    aes(
      x = Size,
      y = Price
    )
  ) +
  geom_bin2d() +
  facet_grid(~ year_2024)
```



2 平均値による要約

2.1 分布の限界

- 変数の数が増えた場合、可視化の手法を用いても、データの分布を示すのは容易ではない
- 多くの応用で、ある変数 Y とその他変数 $X = [X_1, .., X_L]$ の関係性が関心となる
- 例: 価格の特徴把握
  - $Y =$ 価格、 $X =$ [部屋の広さ, 立地]

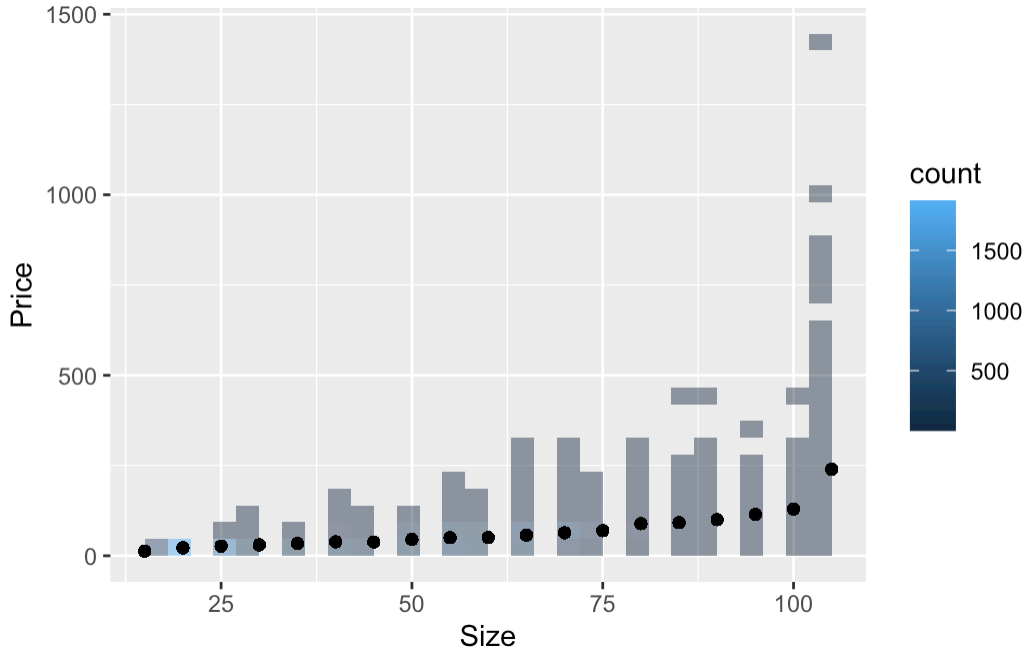
2.2 分布の要約

- Y の分布を、少数の値に要約する
- 代表例: (条件付き)平均値: ある $X = x$ について

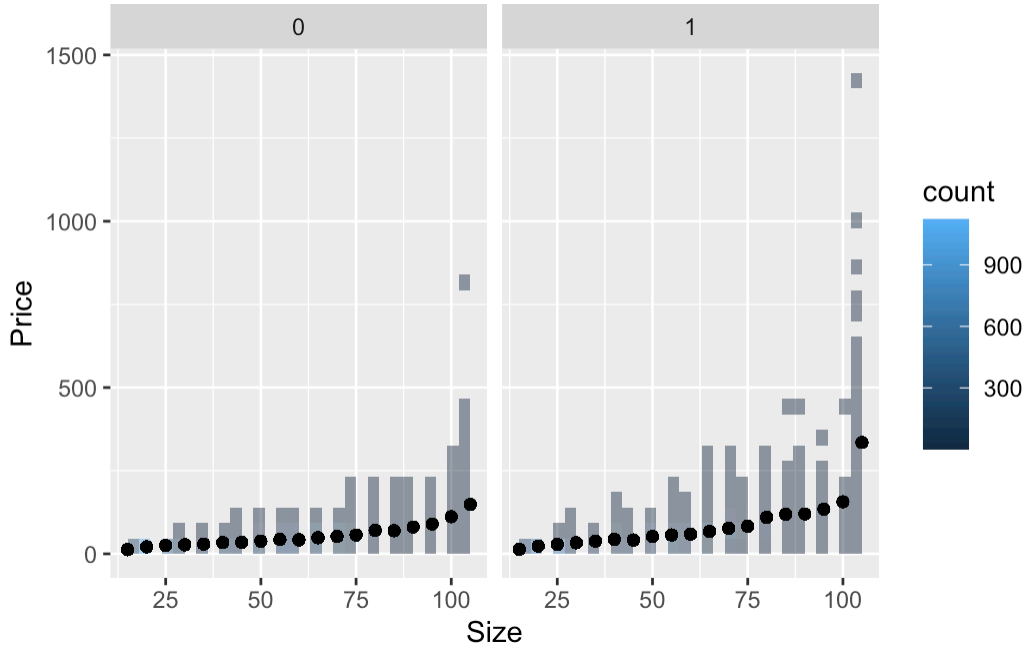
$$\text{平均値} = \frac{Y\text{の総和}}{\text{事例数}}$$

  - 分布情報の多くを捨てていることに注意
  - 例えば、“散らばり度合い”の情報は排除されている
    - 違う指標 (分散など) で捉えることができる

2.3 例: Price ~ Size



2.4 例: Price ~ Size + year_2024



3 線型モデルによる要約

3.1 平均値の限界

- X の数が増えると、平均値の数が膨大に増える
  - 可視化もほぼ不可能

3.2 近似モデル

- モデル化が有効

$$Yの平均値 \simeq モデルの予測値 = \beta_0 + \beta_1 X_1 + .. + \beta_L X_L$$

- β はデータによって決められる値
  - 代表的な決め方は最小二乗法 (OLS)

3.3 最小二乗法

- モデルが計算する値と実際のYとの乖離を最小化するように決める
- 具体的には、以下を最小化

$$\underbrace{(Y - 予測値)^2}_{二乗誤差}の平均値$$

3.4 例

- モデル A $= -200 + 5 \times Size$
- モデル B $= -100 + 10 \times Size$

3.5 例

Price	Size	ModelA	ModelB	二乗誤差(A)	二乗誤差(B)
94	40	0	300	8836	42436
100	65	125	550	625	202500
130	65	125	550	25	176400
98	65	125	550	729	204304
58	40	0	300	3364	58564
330	95	275	850	3025	270400
200	80	200	700	0	250000
430	105	325	950	11025	270400

3.6 最小二乗法の別解釈

- 以下を最小化するように決めても、同じモデルが算出される

$$(Yの平均値 - g(x))^2 \times X = xの割合の総和$$

- 平均値のモデルと解釈できる

3.7 例: Price ~ Size + year_2024

- 以下のモデルを推定

$$Price \simeq g(X) = \beta_0 + \beta_1 \times Size + \beta_2 \times year_{2024}$$

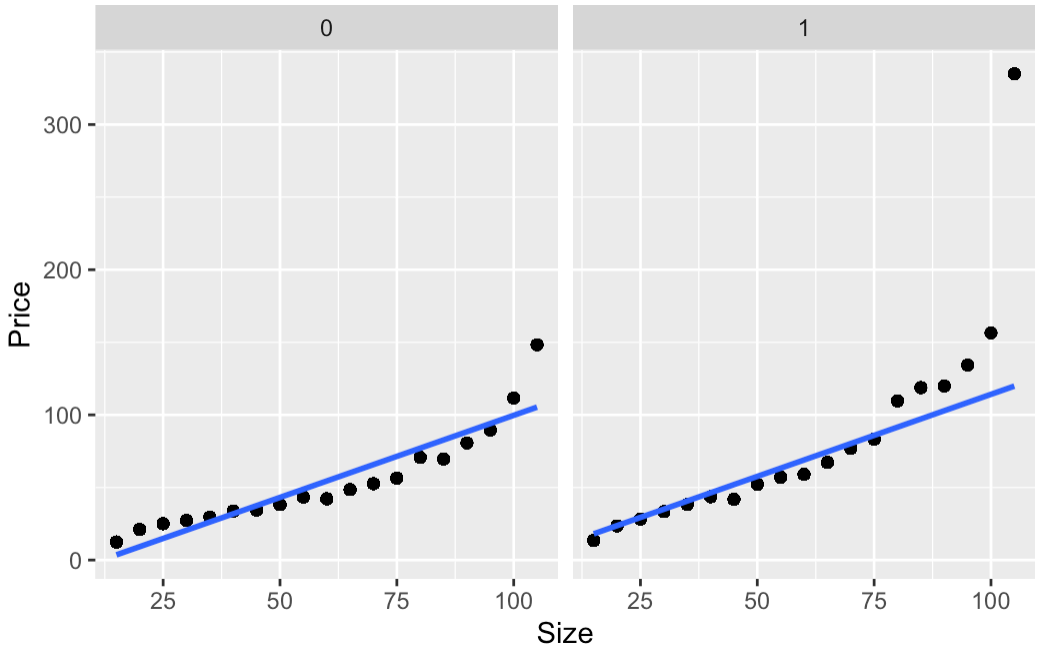
  - year_2024 = 1 (2024 年に取引)/ degree = 0 (2019 年に取引)

3.8 例: Price ~ Size + year_2024

```
lm(Price ~ Size + year_2024, data)
```

```
Call:
lm(formula = Price ~ Size + year_2024, data = data)

Coefficients:
(Intercept)         Size    year_2024
    -13.373        1.131       14.479
```

3.9 例: Price ~ Size + year_2024



3.10 線型モデルの利点

- 可視化が不可能な応用 (X の数が多い)においても、平均値の性質をある程度捉えることができる
  - β の値を見ることで、**近似モデル** において、Y と X の関係性を知ることができる

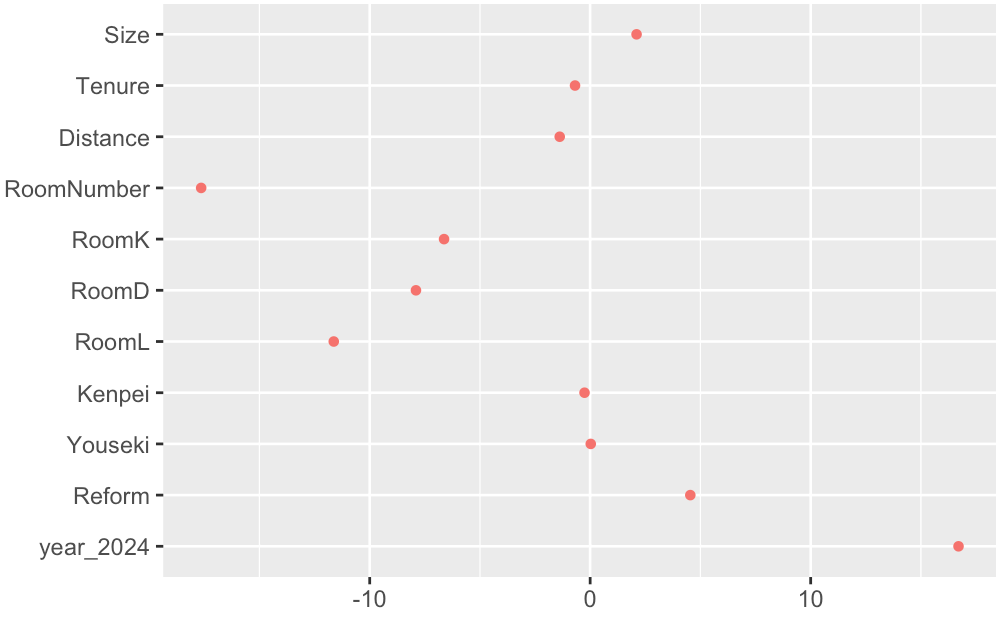
3.11 例

```
model <- lm(Price ~Size + Tenure + Distance + RoomNumber + RoomK + RoomD +
RoomL + Kenpei + Youseki + Reform + year_2024, data)

model
```

```

Call:
lm(formula = Price ~ Size + Tenure + Distance + RoomNumber +
    RoomK + RoomD + RoomL + Kenpei + Youseki + Reform + year_2024,
    data = data)

Coefficients:
(Intercept)         Size       Tenure     Distance   RoomNumber        RoomK
   21.58470      2.10380     -0.68705     -1.38160    -17.64328     -6.62721
      RoomD        RoomL       Kenpei      Youseki       Reform    year_2024
   -7.90176    -11.62979     -0.25619      0.02619      4.54181     16.70609
```

3.12 結果の可視化

```
dotwhisker::dwplot(model, ci = 0)
```



3.13 カテゴリー変数の扱い

- 地域や性別などは、比較的少数のカテゴリーからなる変数も、分析に容易に導入できる
  - ダミー変数に変換すれば OK
- カテゴリー x ダミー: $X = x$ であれば 1、それ以外であれば 0 をとる変数

3.14 例: 立地

```
lm(Price ~ District, data)
```

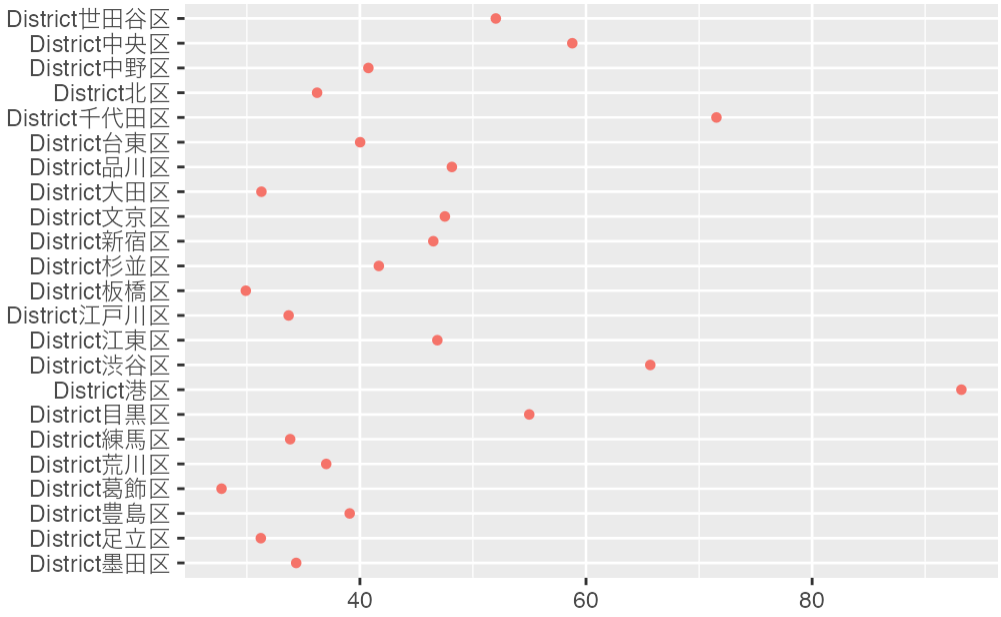
```

Call:
lm(formula = Price ~ District, data = data)

Coefficients:
     (Intercept)    District中央区    District中野区      District北区
          52.021             6.765           -11.274           -15.809
District千代田区    District台東区    District品川区    District大田区
          19.515           -12.001            -3.887           -20.733
  District文京区    District新宿区    District杉並区    District板橋区
          -4.506            -5.530           -10.347           -22.113
District江戸川区    District江東区    District渋谷区      District港区
         -18.326            -5.172            13.663            41.182
  District目黒区    District練馬区    District荒川区    District葛飾区
           2.957           -18.190           -15.002           -24.261
```

```
  District豊島区     District足立区     District墨田区
         -12.926            -20.786            -17.658
```

- カテゴリー(千代田区)のダミー変数は除外される
  - 各値は、千代田と比較した平均取引価格の差

3.15 例: 立地

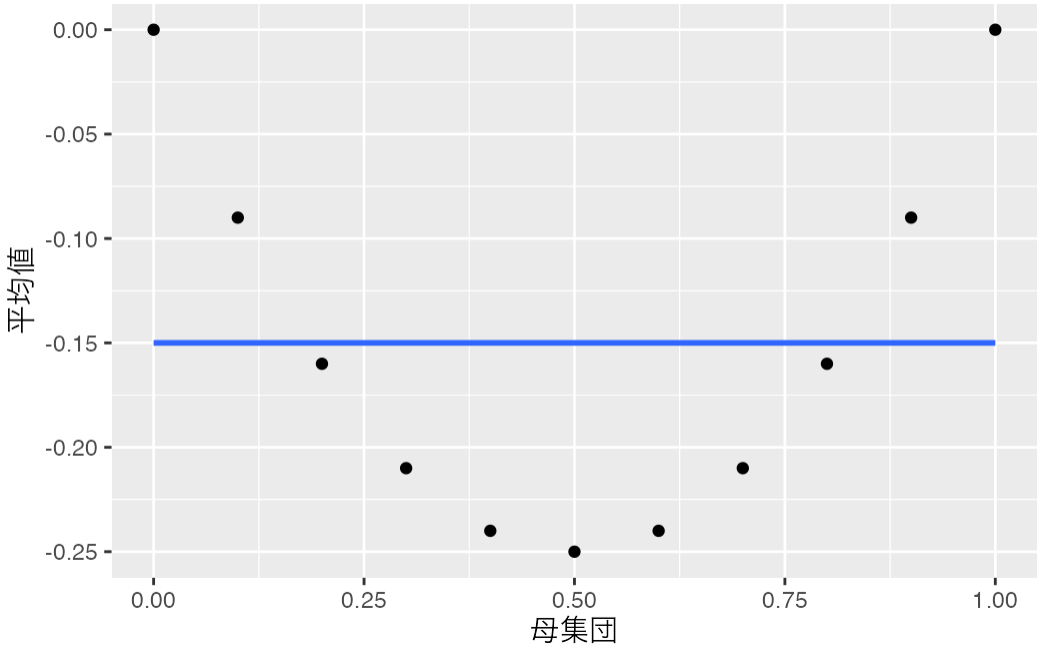
```
model <- lm(Price ~ 0 + District, data) # 定数項を除外する代わりに、千代田区ダミーを導入

dotwhisker::dwplot(model, ci = 0)
```



3.16 線型モデルの限界

- あくまでも近似モデルであり、平均値 と モデルの特徴は大きく乖離しうる

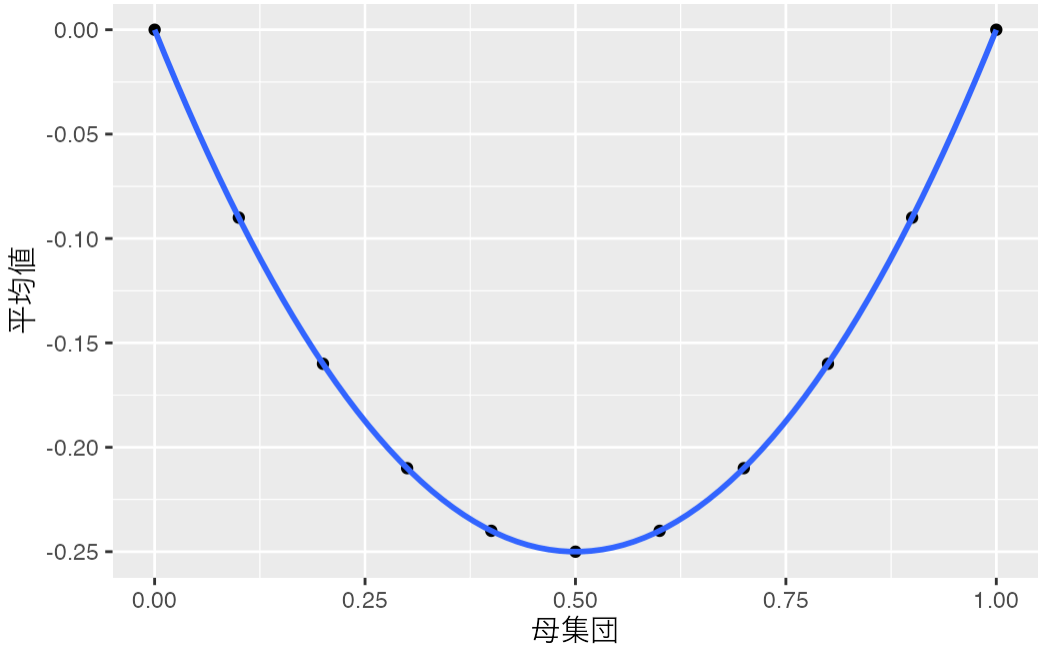
3.17 例



3.18 モデルの複雑化

- より複雑なモデルを推定することで、データへの適合度を改善できる
  - 誤定式を減らせる
  - どこまで複雑化すべきかは、研究目標によって異なる

3.19 例. 複雑なモデル



3.20 Takeaway

- 分布 → 平均 → 近似モデル の順番に、情報量を減らすことを受け入れながら、人間が理解しやすくしている
  - 分布 ≠ 平均 ≠ 近似モデル を常に意識することが重要
- 可能な限り、近似モデルと散布図や平均値を同時に図示することを推奨

Bibliography