第 3 章: 社会調査研究による母集団の特徴の推論

Elena Llaudet & Kosuke Imai.

Data Analysis for Social Science: A Friendly and Practical Introduction.

2026-03-09

3.1 英国における EU 国民投票

2016 年 Brexit 国民投票

- 2016 年、英国 (UK) で EU 離脱の是非を問う国民投票が実施された（Brexit）。
- 英国選挙調査 (BES) は、有権者の意識を調査するために大規模な社会調査を実施。
- **目標**: 母集団（全有権者）の離脱支持率を推定し、支持者の特徴（学歴、年齢など）を明らかにする。

3.2 社会調査研究

標本と母集団

- **母集団** (Population): 調査対象となる集団全体。観察数を N で表す。
- **標本** (Sample): 母集団から選ばれた個人の部分集合. 観察数を n で表す。
- **代表的な標本** (Representative sample): 母集団の特徴を正確に反映している標本。
- **無作為抽出** (Random sampling): 母集団から無作為に標本を選ぶことで、平均的に母集団を代表することを保証する。

潜在的な問題点

1. **抽出枠の不備**: 母集団のリスト（抽出枠）が不完全。
2. **全項目無回答**: 調査参加の拒否。
3. **一部項目無回答**: 特定の質問への回答拒否。
4. **誤った申告**: 社会的に望ましい回答をするなどの虚偽報告。

3.3 Brexit への支持の測定

BES データの読み込み (1)

- 有権者の意識調査データを R に読み込みます。

```
# 1. ローカルに保存した BES データの読み込み（推奨）
bes <- read.csv("BES.csv")

#（参考）URL から直接読み込むことも可能
# bes <- read.csv("https://ayumu-tanaka.github.io/QSS/DSS_Data/BES.csv")
```

データの確認 (2)

```
# 1. データの最初の数行を表示
head(bes)
```

```
##          vote leave education age
## 1       leave     1         3  60
## 2       leave     1        NA  56
## 3        stay     0         5  73
## 4       leave     1         4  64
## 5 don't know    NA         2  68
## 6        stay     0         4  85
```

```
# 2. 標本サイズの確認
dim(bes)
```

```
## [1] 30895     4
```

度数表と比率表

- table() で各カテゴリの回答数をカウントする。
- prop.table() で比率（割合）に変換する。

```
# 回答の度数表
table(bes$vote)
```

```
##
## don't know      leave       stay won't vote
##       2314      13692      14352        537
```

```
# 回答の比率表
prop.table(table(bes$vote))
```

```
##
## don't know      leave       stay won't vote
## 0.07489885 0.44317851 0.46454119 0.01738145
```

3.4 Brexit を支持したのは誰？

欠損データの処理 (1)

- R では欠損値を NA で表す。

```
# 欠損値を含む教育水準の確認
table(bes$education, exclude = NULL)
```

```
##
##     1     2     3     4     5  <NA>
##  2045  5781  6272 10676  2696  3425
```

欠損データの処理 (2)

- `mean()` などの関数では、`na.rm = TRUE` を指定して NA を除外する必要がある。
- `na.omit()` は、NA を含む行をすべて削除する。

```
# NA を除外して平均を計算
mean(bes$leave, na.rm = TRUE)
```

```
## [1] 0.4882328
```

```
# NA を含む行を削除した新しいデータフレーム作成
bes1 <- na.omit(bes)
```

二元度数表

- ２つの変数の関係を「クロス集計表」で確認する。

```
# 離脱支持 (leave) と教育水準 (education) の二元度数表
table(bes1$leave, bes1$education)
```

```
##
##         1    2    3    4    5
##    0  498 1763 3014 6081 1898
##    1 1356 3388 2685 3783  631
```

二元比率表

- 行ごとの比率を確認する。

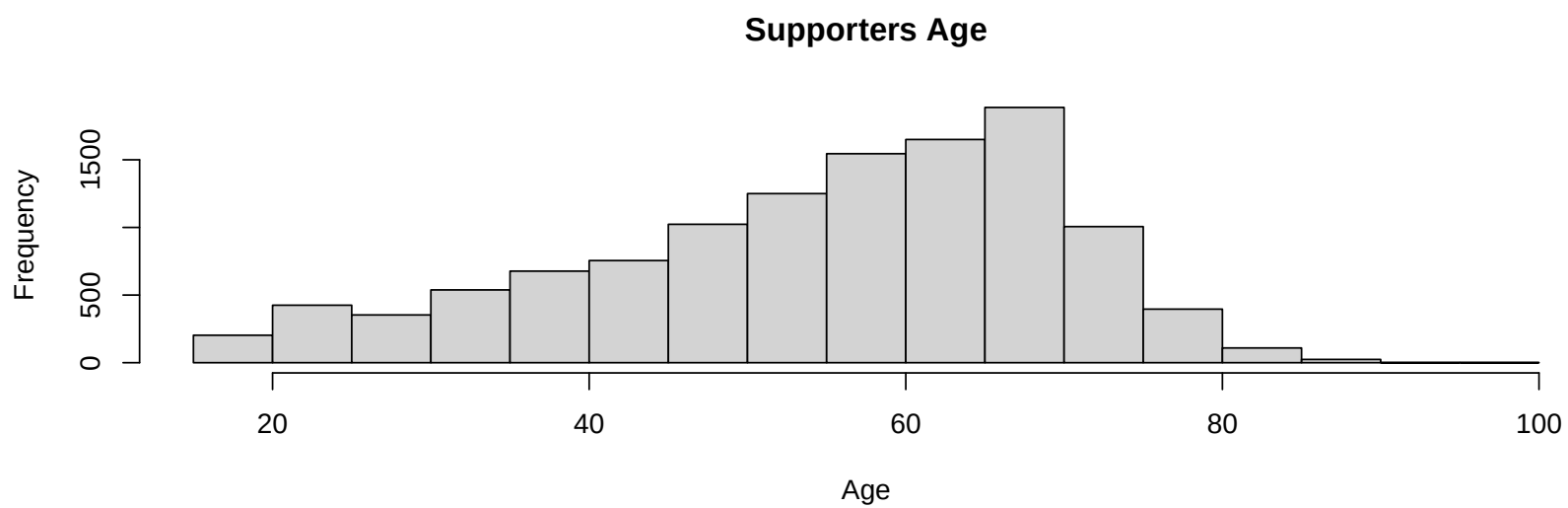
```
# 行ごとの比率 (margin = 1)
prop.table(table(bes1$leave, bes1$education), margin = 1)
```

```
##
##                1           2           3           4           5
##   0 0.03757356 0.13301645 0.22740305 0.45880489 0.14320205
##   1 0.11449802 0.28607616 0.22671620 0.31942920 0.05328042
```

ヒストグラム（度数と密度）

- `hist()` で数値変数の分布を視覚化する。
- `freq = FALSE` で密度ヒストグラム（面積の合計が 1）を作成できる。

```
# 年齢のヒストグラム（離脱支持者のみ）
hist(bes1$age[bes1$leave == 1],
     main = "Supporters Age", xlab = "Age")
```



記述統計量

- 分布の中心（平均値、中央値）とばらつき（標準偏差、分散）を計算する。

```
# 離脱支持者の平均年齢
mean(bes1$age[bes1$leave == 1])
```

```
## [1] 55.06823
```

```
# 離脱支持者の年齢の中央値
median(bes1$age[bes1$leave == 1])
```

```
## [1] 58
```

```
# 標準偏差
sd(bes1$age[bes1$leave == 1])
```

```
## [1] 14.96106
```

3.5 教育と離脱派票との関係

UK_districts データの読み込み (1)

- 地区レベルのデータを読み込み、欠損値を除外します。

```
# 1. ローカルに保存したデータの読み込み（推奨）
dis <- read.csv("UK_districts.csv")

#（参考）URL から直接読み込むことも可能
# dis <- read.csv("https://ayumu-tanaka.github.io/QSS/DSS_Data/UK_districts.csv"

# 2. NA を除外
dis1 <- na.omit(dis)
```

データの構造確認 (2)

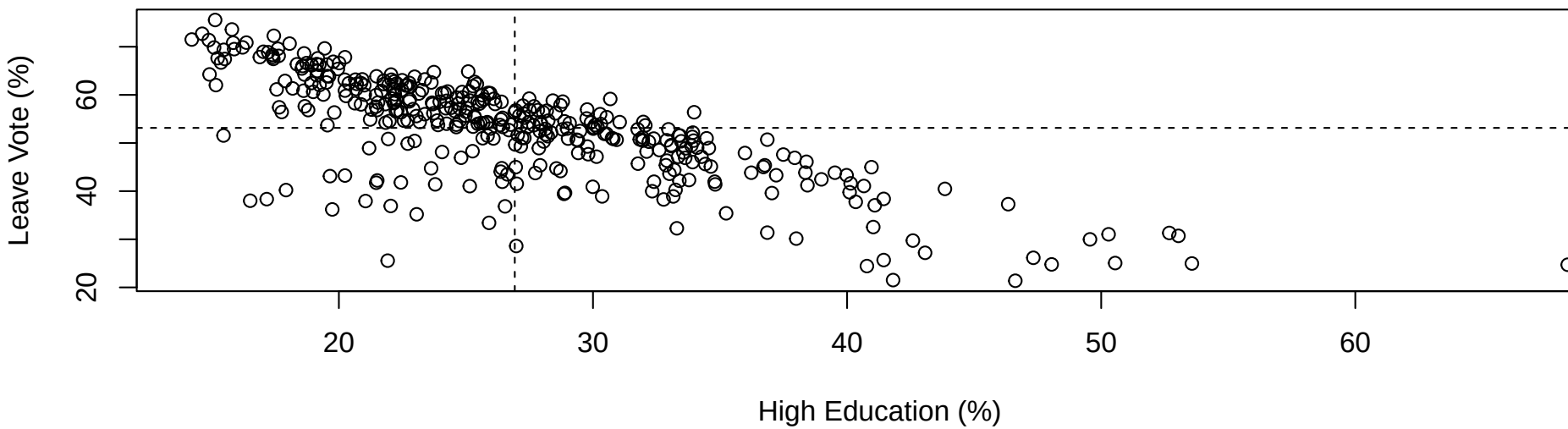
```
# 最初の数行を表示
head(dis1)
```

```
##              name leave high_education
## 1      Birmingham 50.42           22.98
## 2         Cardiff 39.98           32.33
## 3 Edinburgh City 25.56           21.92
## 4   Glasgow City 33.41           25.91
## 5       Liverpool 41.81           22.44
## 6         Swansea 51.51           25.85
```

散布図

- plot() で 2 つの数値変数の関係をプロットする。
- abline() で平均値を示す直線などを追加できる。

```
plot(dis1$high_education, dis1$leave,
     xlab = "High Education (%)",
     ylab = "Leave Vote (%)")
abline(v = mean(dis1$high_education), lty = "dashed")
abline(h = mean(dis1$leave), lty = "dashed")
```



相関

- cor() で 2 つの変数の間の線形関係の強さと方向を数値化する（-1 から 1 の間）。

```
# 高学歴率と離脱派得票率の相関
cor(dis1$high_education, dis1$leave)
```

```
## [1] -0.7633185
```

- **解釈**: 強い負の相関がある。教育水準が高い地区ほど、離脱派の得票率が低い傾向にある。

3.6 まとめ

第3章のまとめ

- **測定**: 母集団を代表する標本を用いて、関心のある数量を推定する。
- **無作為抽出**: 推論の妥当性を高めるための基本的な手法。
- **データの要約と視覚化**:
  - 度数表、比率表、クロス表。
  - ヒストグラム、散布図。
  - 平均、中央値、標準偏差、相関係数。
- **R の操作**: `table()`, `prop.table()`, `na.omit()`, `hist()`, `plot()`, `cor()` を学んだ。