第 6 章: 確率

Elena Llaudet & Kosuke Imai.

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

2026-03-09

6.1 確率の重要性

なぜ確率を学ぶのか？

- データ分析には常に**不確実性** (Uncertainty) が伴う。
- 標本から母集団を推測する際、その推測がどれくらい正確かを評価するために確率の知識が必要。
- 社会科学における例：
  - 世論調査の誤差の評価
  - 政策効果が偶然ではないことの確認
  - 将来の出来事の発生予測

6.2 ベルヌーイ分布

ベルヌーイ分布 (Bernoulli Distribution)

- 2 つの結果（成功/失敗、1/0）しかない試行の確率分布。
- 例：コイン投げ（表=1, 裏=0）

```
# 100 万回のコイン投げシミュレーション
possible_values <- c(1, 0)
flips <- sample(possible_values, size = 1000000,
                replace = TRUE, prob = c(0.5, 0.5))

# 比率の確認
prop.table(table(flips))
```

```
## flips
##        0        1
## 0.500048 0.499952
```

```
# 平均（期待値）
mean(flips)
```

```
## [1] 0.499952
```

6.3 正規分布

STAR データの読み込み (1)

- 連続データの分布（正規分布）を確認するため、データを読み込みます。

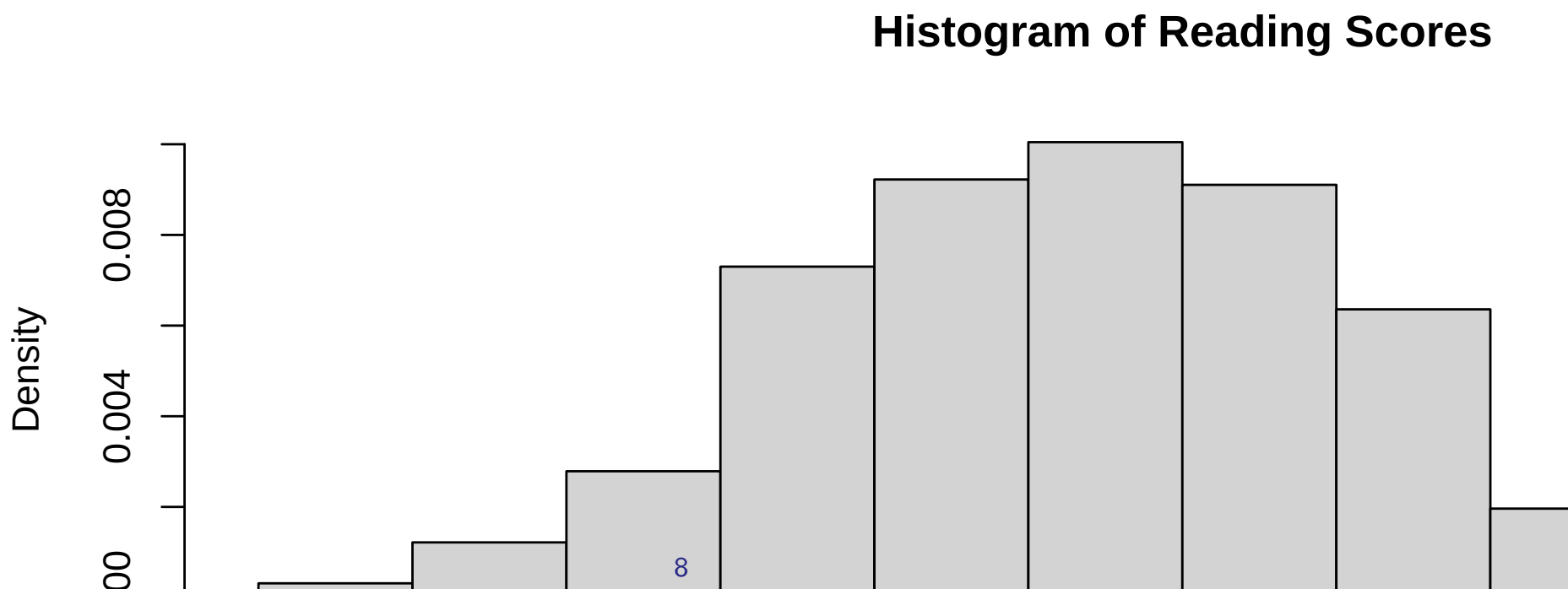
```
# 1. ローカルに保存したデータの読み込み（推奨）
star <- read.csv("STAR.csv")

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

正規分布 (Normal Distribution) (2)

- 左右対称の釣鐘型の分布。多くの自然・社会現象に現れる。
- 平均 (μ) と標準偏差 (σ) で形状が決まる。

```
# 読解力テストの得点分布
hist(star$reading, freq = FALSE,
     main = "Histogram of Reading Scores",
     xlab = "Reading Score")
```



標準正規分布 (Standard Normal Distribution)

- 平均 = 0, 標準偏差 = 1 の正規分布。
- pnorm() 関数を用いて、ある値以下の確率を計算できる。

```
# Z <= -1.96 となる確率（約 2.5%）
pnorm(-1.96)
```

```
## [1] 0.0249979
```

```
# -1.96 <= Z <= 1.96 となる確率（約 95%）
pnorm(1.96) - pnorm(-1.96)
```

```
## [1] 0.9500042
```

6.4 大数の法則

大数の法則 (Law of Large Numbers)

- サンプルサイズ (n) が大きくなるほど、標本平均は母平均に近づく。

```
# 母集団の支持率 p = 0.6 とする
# n = 10 の場合
mean(sample(c(1, 0), size = 10, replace = TRUE,
            prob = c(0.6, 0.4)))
```

```
## [1] 0.6
```

```
# n = 1000 の場合
mean(sample(c(1, 0), size = 1000, replace = TRUE,
            prob = c(0.6, 0.4)))
```

```
## [1] 0.601
```

```
# n = 1000000 の場合
mean(sample(c(1, 0), size = 1000000, replace = TRUE,
            prob = c(0.6, 0.4)))
```

```
## [1] 0.599924
```

6.5 中心極限定理

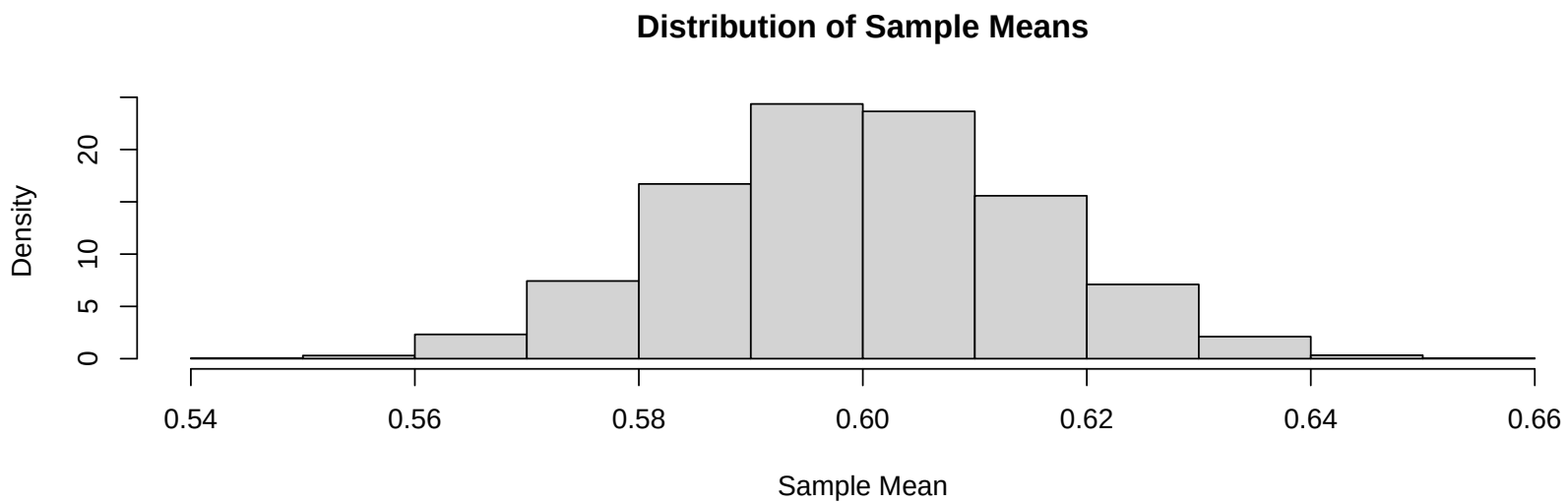
中心極限定理 (Central Limit Theorem)

- もとの分布が何であれ、サンプルサイズが十分に大きければ、標本平均の分布は**正規分布**に近づく。
- 推測統計学の最も重要な基盤。

```
# シミュレーション：標本平均の分布
sample_means <- c()
for(i in 1:10000){
  sample_means[i] <- mean(sample(c(1, 0),
                                 size = 1000,
         replace = TRUE, prob = c(0.6, 0.4)))
}
```

中心極限定理の視覚化

```
hist(sample_means, freq = FALSE,
     main = "Distribution of Sample Means",
     xlab = "Sample Mean")
```



6.6 まとめ

第6章のまとめ

- **確率**は不確実性を数値化するための道具である。
- **ベルヌーイ分布**は二値データを、**正規分布**は連続データをモデル化する。
- **大数の法則**により、大きなサンプルはより正確な推測を可能にする。
- **中心極限定理**により、標本平均の挙動を予測し、信頼区間や仮説検定に繋げることができる（次章）。