6. 母集団と標本をシミュレーションで理解する

honocat

2025-12-18

母集団と標本のシミュレーション

母集団を用意する

例として、女性 4,600 人、男性 5,400 人から成る母集団を考える。合計 10,000 人で、女性比率が 0.46、男性比率は 0.54 である。これらは母集団の比率なので、母比率と呼ばれる。

この母集団(population)を R で定義する。

```
pop <- rep(c('female', 'male'), c(4600, 5400))
length(pop) # 総人口
```

```
[1] 10000
```

```
table(pop) # 男女の数
```

```
pop
female   male
  4600   5400
```

ここで、私たちは母比率を知らないと仮定する。正しい母比率を調べるもっとも単純な方法は、1 万人全員の性別を調べることである。しかし、1 万人を調査するのは大変なので、1 万人から 100 人だけを無作為(ランダム)に選び、100 人の性別を調べ、その結果を利用して母比率を推定することにする。

母集団から 100 人をランダムに選ぶ

母集団から、ランダムに 100 人を抜き出してみる。`sample()` を使う。

```
N <- 100
sample_1 <- sample(pop, size = N, replace = FALSE)
table(sample_1) / N
```

```
sample_1
female   male
  0.42   0.58
```

この比率は標本(サンプル)の比率なので、標本比率と呼ばれる。もう一度調べてみる。

```
sample_2 <- sample(pop, size = N, replace = FALSE)
mean(sample_2 == 'female')
```

```
[1] 0.4
```

もう一度。

```
sample_3 <- sample(pop, size = N, replace = FALSE)
mean(sample_3 == 'female')
```

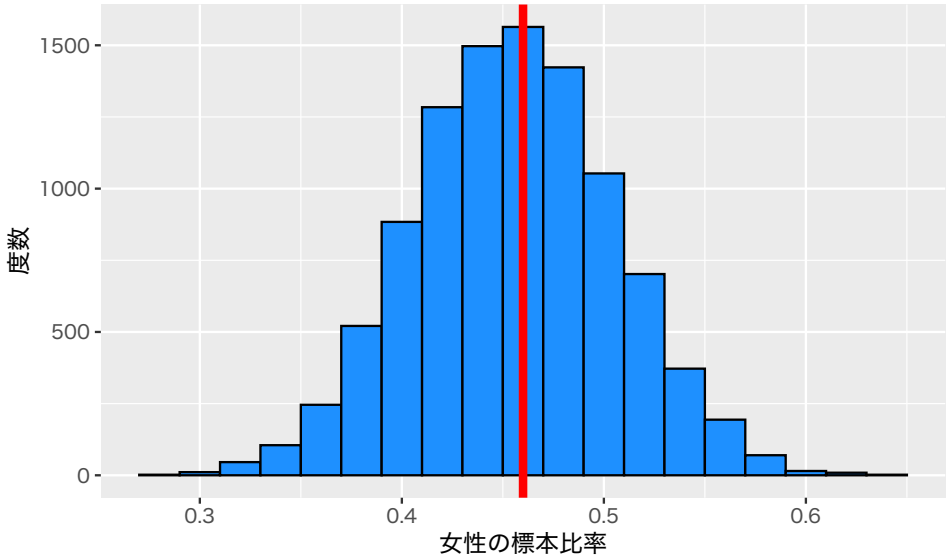
```
[1] 0.47
```

1 万人から 100 人を選ぶ方法は 6.5×10^{241} 通りあるので、すべての組み合わせを調べるのは不可能。R を使って 10,000 通りだけ調べる。

```
res_1 <- rep(NA, 1e4)
for (i in 1 : 1e4) {
  x <- sample(pop, size = N, replace = FALSE)
  res_1[i] <- mean(x == 'female')
}
```

結果をヒストグラムにしてみる。私たちは母比率を知っているので、母比率である 0.46 を赤い線で示す。

```
df1 <- tibble(sample = res_1)
hist1 <- ggplot(df1, aes(x = sample)) +
  geom_histogram(binwidth = 0.02,
                 color    = 'black',
                 fill     = 'dodgerblue') +
  geom_vline(xintercept = 0.46,
             color      = 'red',
             linewidth  = 1.5) +
  labs(x = '女性の標本比率',
       y = '度数')
plot(hist1)
```



ヒストグラムを見ると、一つ一つの標本比率は母比率よりも大きかったり、母比率よりも小さかったりする。しかし、**平均すると**母比率に近い値をえることができそう。

このヒストグラムから分かる通り、統計量は分布する（つまり、標本ごとにばらばらの値を取る）。このような標本ごとの分布を**標本分布（sampling distribution）**と呼ぶ。

標準誤差（standard error; SE）

標本分布に現れる標準偏差（統計量のばらつき）を標準誤差という。このシミュレーションで得られる標準誤差は、

```
sd(res_1)
```

```
[1] 0.04961149
```

である。理論的には、

$$\mathrm{SE} = \frac{\text{母標準偏差}}{\sqrt{\text{標本サイズ}}}$$

なので、

```
pi <- 0.46
pop_sd <- sqrt(pi * (1 - pi))
pop_sd / sqrt(100)
```

```
[1] 0.04983974
```

になるはずであるが、シミュレーションなので理論値に近づいたり離れたりする。