9. 母平均の推定

honocat

2025-12-20

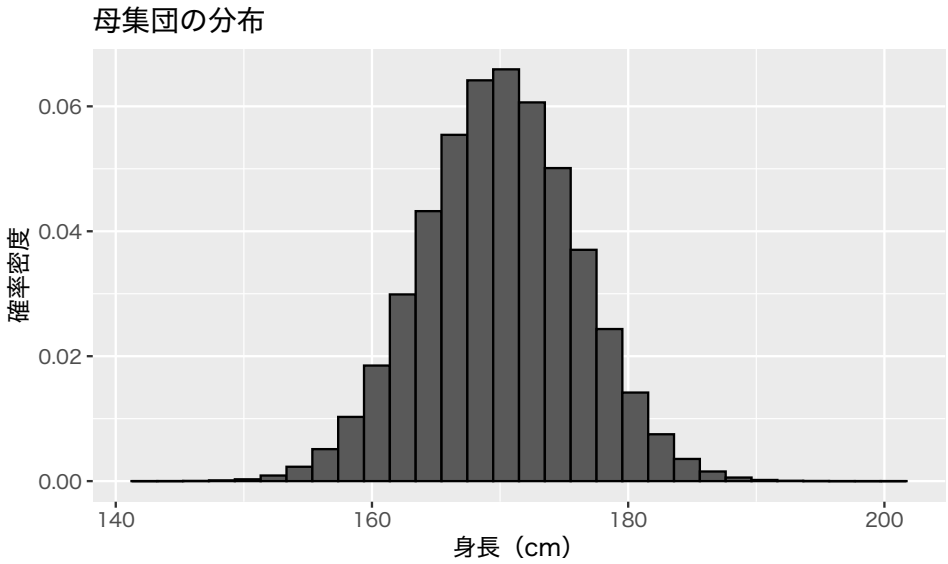
母集団を定義する

成人男性の身長に興味があるとする。母集団の人口を 100 万人、母平均を約 170cm、母標準偏差を約 6cm に設定する。

```
pop <- rnorm(1e6, mean = 170, sd = 6)
```

母集団の身長の分布は以下。

```
pop_height <- ggplot(tibble(pop),
                     aes(x = pop,
                         y = after_stat(density))) +
  geom_histogram(color = 'black') +
  labs(x = '身長(cm)',
       y = '確率密度',
       title = '母集団の分布')
plot(pop_height)
```



この母集団における統計量は、

```
mean(pop)
```

```
[1] 170.0021
```

```
sd(pop)
```

```
[1] 6.002633
```

である。

標本を抽出して母平均を推定する

標本 1

100 万人の母集団全員を調べるのではなく、100 人だけ標本として抽出し、その標本を利用して母平均を推定してみる。

標本サイズ 100 の標本を 1 つ抽出する。

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

■点推定（point estimation）

標本平均を計算する

```
mean(sample_1)
```

```
[1] 169.9813
```

これが母平均の点推定値(point estimate)である。

■区間推定(inteval estimation)

区間推定では、1 つの値を示す代わりに推定に区間を利用することで、推定に対する不確実性を示す。

推定に標準正規分布を利用する場合、身長 h の点推定値を $\bar{h}$ とすると、以下のように定義される信頼区間(confidence interval)を区間推定に使う。

$$[\bar{h} - Q \cdot \mathrm{SE}, \bar{h} + Q \cdot \mathrm{SE}]$$

ここで、SE は標準誤差(standard error)で、これは以下のように推定する。

$$\mathrm{SE} = \frac{u}{\sqrt{N}}$$

N は標本サイズ、u は不偏分散の平方根。

また、Q (と $-Q$)はどのような信頼区間を求めたいかによって変わる。例えば、95% 信頼区間を求めたいときは、$Q = 1.96$ を使う。

特定の信頼度(信頼度に何 % を使うか)に対する Q の求め方は以下の通り。まず、100% から信頼度を引く。95% の場合 $1 - 0.95 = 0.05$ である。次に、その値を 2 で割る。95% の場合、$\frac{0.05}{2} = 0.025$ である。この値を `qnorm()` に当てはめる。ただし、`lower.tail = FALSE` を指定する。

```
(Q_95 <- qnorm((1 - 0.95) / 2, lower.tail = FALSE))
```

```
[1] 1.959964
```

少数第 2 位までで丸めると、

```
round(Q_95, digit = 2)
```

```
[1] 1.96
```

である。次に SE を求めよう。そのためにまず、標本の分散(不偏分散, unbiased variance)を計算する。身長を h とすると、身長の不偏分散 $\mathrm{Var}(h)$ は、

$$\mathrm{Var}(h) = \frac{\sum_{i=1}^{n} (h_i - \bar{h})^2}{N - 1}$$

と定義される。これは、var() で計算できる。

```
(var_1 <- var(sample_1))
```

```
[1] 35.09249
```

この平方根が、不偏分散の平方根(u)である。

```
(sd_1 <- sqrt(var_1))
```

```
[1] 5.923891
```

この値は sd() でも求められる。

```
sd(sample_1)
```

```
[1] 5.923891
```

母分散または母標準偏差を知らないときは、ここで計算した標本から得られる値を推定値として使う。ここでは、標本の標準偏差(不偏分散の平方根)を母標準偏差の推定値として使う。このサンプルから推定される SE、

```
(SE_1 <- sd_1 / sqrt(N))
```

```
[1] 0.5923891
```

となる。以上から、95% 信頼区間の上下限は、

```
(lb <- mean(sample_1) - Q_95 * SE_1)
```

```
[1] 168.8202
```

```
(ub <- mean(sample_1) + Q_95 * SE_1)
```

```
[1] 171.1423
```

よって、この標本から得られる 95% 信頼区間は、[168.82, 171.14] である。

信頼区間を求める関数を作る

```
get_confint <- function(x, level = 0.95) {
  ## 標準正規分布を利用して信頼関数を求める関数
  ## 引数: x = 推定に使う標本
  ##       level = 信頼度。期待値は 0.95
  ## 返り値: 点推定値、信頼区間の上下限の 3 つを含むベクトル
  N <- length(x)
  mean_x <- mean(x)
  SE <- sd(x) / sqrt(N)
  Q <- qnorm((1 - level) / 2, lower.tail = FALSE)
  lb <- mean_x - Q * SE
  ub <- mean_x + Q * SE
  estimates <- c(round(mean_x, 2), round(lb, 2), round(ub, 2))
  names(estimates) <- c('点推定値',
                        str_c(100 * level, '%CI の下限'),
                        str_c(100 * level, '%CI の上限'))
  return(estimates)
}
```

正しく動くか確認。

```
get_confint(sample_1)
```

```
    点推定値 95%CI の下限 95%CI の上限
      169.98       168.82       171.14
```

標本 2

別の標本を用意する。

```
sample_2 <- sample(pop, size = N, replace = FALSE)
get_confint(sample_2)
```

```
    点推定値 95%CI の下限 95%CI の上限
      171.24       169.97       172.51
```

標本 3

別の標本を用意する。

```
sample_3 <- sample(pop, size = N, replace = FALSE)
get_confint(sample_3)
```

```
   点推定値 95%CI の下限 95%CI の上限
     169.61       168.42       170.81
```

シミュレーション

標本抽出を何度も繰り返し、推定(の誤差)がどのようにばらつくか確認する。

シミュレーションは 1 万回とする。

```
n_sims <- 1e4
```

結果を保存する容器を用意する。

```
res_sim <- matrix(NA, nrow = n_sims, ncol = 3)
colnames(res_sim) <- c('est', 'lb', 'ub')
```

for ループでシミュレーションを行う。

```
for (i in 1 : n_sims) {
  smpl <- sample(pop, size = N, replace = FALSE)
  res_sim[i,] <- get_confint(smpl)
}
res_sim[1 : 5,]
```

```
        est     lb     ub
[1,] 168.88 167.72 170.05
[2,] 168.96 167.79 170.12
[3,] 169.84 168.59 171.09
[4,] 170.65 169.34 171.96
[5,] 171.34 170.23 172.45
```

不編成を確認する

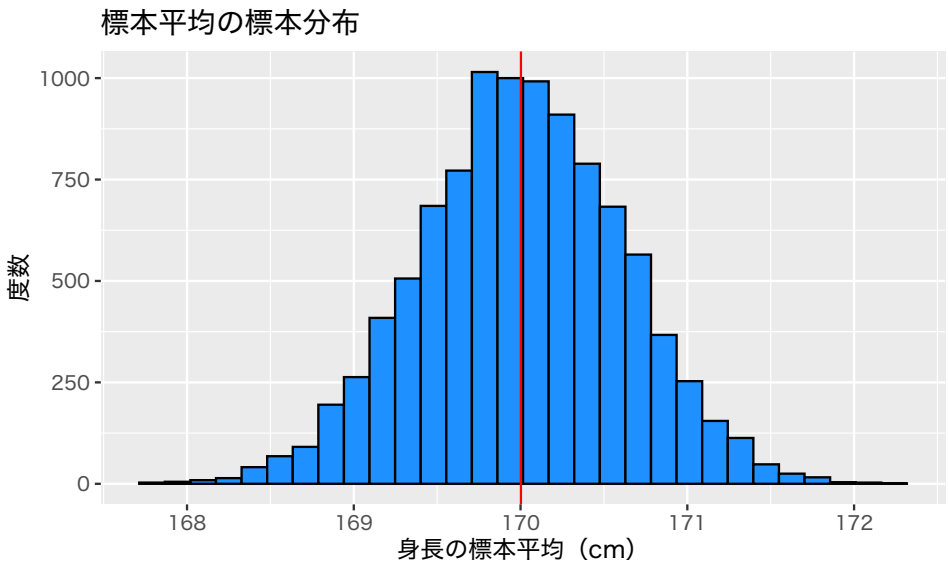
シミュレーションの結果を使って、点推定値のヒストグラムを作ってみる。

```
df_sim <- as_tibble(res_sim)
glimpse(df_sim)
```

```
Rows: 10,000
Columns: 3
$ est <dbl> 168.88, 168.96, 169.84, 170.65, 171.34, 170.26, 170.72, 170.74, 17~
$ lb  <dbl> 167.72, 167.79, 168.59, 169.34, 170.23, 169.01, 169.54, 169.58, 17~
$ ub  <dbl> 170.05, 170.12, 171.09, 171.96, 172.45, 171.51, 171.89, 171.89, 17~
```

```
df_sim$id <- 1 : n_sims
```

```
h_est <- ggplot(df_sim, aes(x = est)) +
  geom_histogram(color = 'black',
                 fill  = 'dodgerblue') +
  geom_vline(xintercept = mean(pop),
             color      = 'red') +
  labs(x = '身長の標本平均(cm) ',
       y = '度数',
       title = '標本平均の標本分布')
plot(h_est)
```



これらの**標本平均の平均は母平均に一致する**というのが「不偏性」である。

```r
mean(df_sim$est)
```

```
[1] 170.0041
```

```r
mean(pop)
```

```
[1] 170.0021
```

標本平均の平均と母平均がほぼ一致することがわかる。

信頼区間の意味を理解する

ひとつの信頼区間を取り出してみる。

```r
res_sim[100, 2 : 3]
```

```
    lb     ub
168.76 171.19
```

この信頼区間に、真の母平均 170.0020841 は含まれているだろうか。母平均が区間内にあれば、この 95% 信頼区間が母数（パラメタ）を区間内に捉えている確率は 1（100%）である。反対に、母数がこの区間内になければ、この 95% 信頼区間が母数を区間内にとらえている確率は 0 である。

これを、1 つひとつの標本について確かめ、1 万回行ったシミュレーションのうち、母数を捉える 95% 信頼区間がいくつ得られたか数えてみる。

```r
caught <- res_sim[, 2] <= mean(pop) & mean(pop) <= res_sim[, 3]
```

条件を 2 つとも満たす場合は TRUE、そうでない場合は FALSE になる。

```r
caught[1 : 5]
```

```
[1]  TRUE  TRUE  TRUE  TRUE FALSE
```

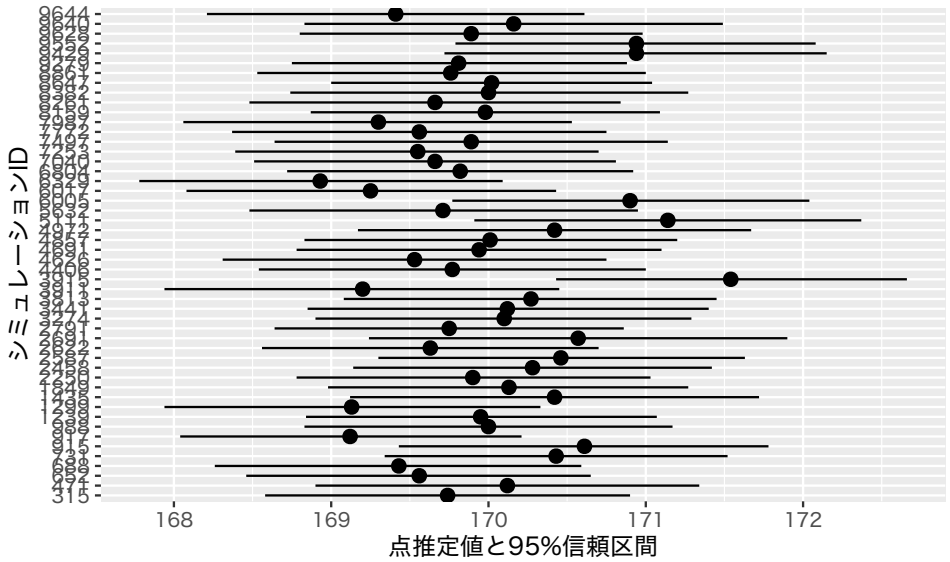
```r
sum(caught)
```

```
[1] 9464
```

である。つまり、1 万個の 95% 信頼区間のうち、母数を捉えることができた区間は 9464 個である。言い換えると、標本抽出と区間推定を繰り返し行うと、得られた 95% 信頼区間の 94.64（約 95%）は、母数を区間内に捉えられる。これが信頼区間の意味である。

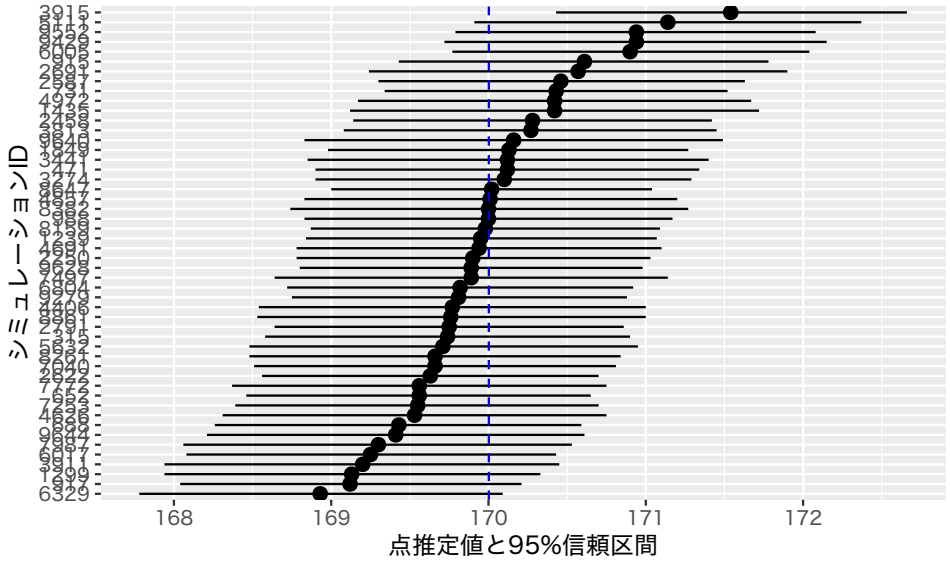
図示してみる。信頼区間を無作為に 50 個選ぶ。

```
rdm_50 <- slice_sample(df_sim, n = 50)
ci1 <- ggplot(rdm_50,
              aes(y = as.factor(id),
                  x = est,
                  xmin = lb,
                  xmax = ub)) +
  geom_pointrange() +
  labs(y = 'シミュレーション ID',
       x = '点推定値と 95% 信頼区間')
plot(ci1)
```



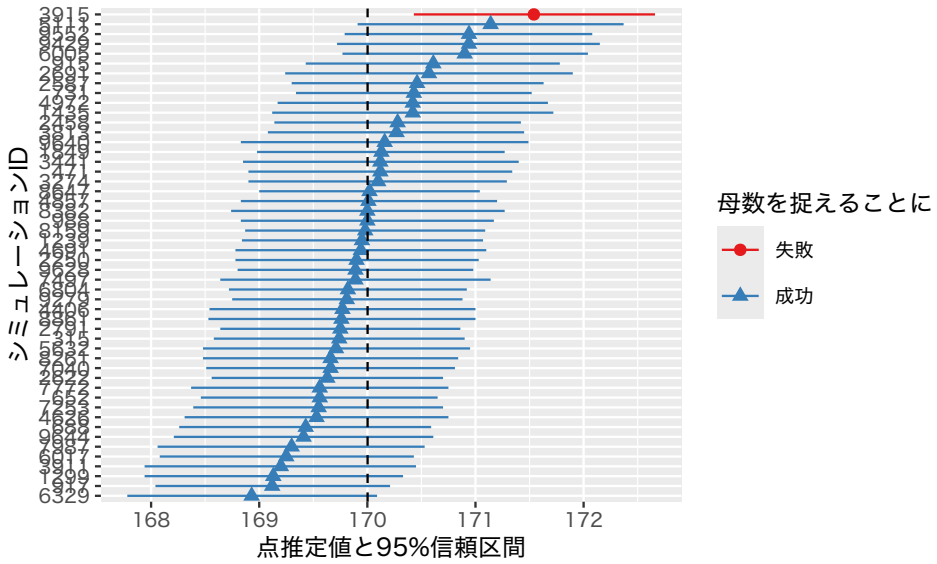
結果を見やすくするため、点推定値の大きさで並べ替える。

```
ci2 <- ggplot(rdm_50, aes(y = as.factor(reorder(id, est)),
                          x = est,
                          xmin = lb,
                          xmax = ub)) +
  geom_pointrange() +
  geom_vline(xintercept = mean(pop),
             color      = 'blue',
             linetype   = 'dashed') +
  labs(y = 'シミュレーション ID',
       x = '点推定値と 95% 信頼区間')
plot(ci2)
```



区間が母数を捉えているかどうかで色を変える。

```
rdm_50$caught <- rdm_50$lb <= mean(pop) & mean(pop) <= rdm_50$ub
le_name <- '母数を捉えることに'
le_labels <- c('失敗', '成功')
ci3 <- ggplot(rdm_50,
              aes(y = as.factor(reorder(id, est)),
                  x = est,
                  xmin = lb,
                  xmax = ub,
                  color = caught,
                  shape = caught)) +
  geom_pointrange() +
  geom_vline(xintercept = mean(pop),
             color      = 'black',
             linetype   = 'dashed') +
  labs(y = 'シミュレーション ID',
       x = '点推定値と 95% 信頼区間') +
  scale_color_brewer(palette = 'Set1',
                     name    = le_name,
                     labels  = le_labels) +
  scale_shape_discrete(name = le_name,
                       labels = le_labels)
plot(ci3)
```



50 個の 95% 信頼区間のうち、1 個の信頼区間が母数を捉えそこねている。