# 8. 標本分布を理解する

honocat

2025-12-19

## 標本分布のシミュレーション

### シミュレーション問題の設定

母集団(成人男性全体)では、平均身長が 170cm、身長の標準偏差が 5.5cm であり、身長は正規分布に従うことを知っているとする。

```
mu <- 170
sigma <- 5.5
```

様々な標本サイズで標本を抽出し、その標本の平均身長を求める作業を 1 万回ずつ繰り返す。

```
n_sims <- 1e4
```

### 標本サイズ($N$)が 1 のとき

標本サイズ $N = 1$ で標本を抽出し、標本平均を計算することを 10,000 回繰り返す。

```
N <- 1
h <- rnorm(N, mean = mu, sd = sigma)
mean(h)
```

```
[1] 171.0474
```

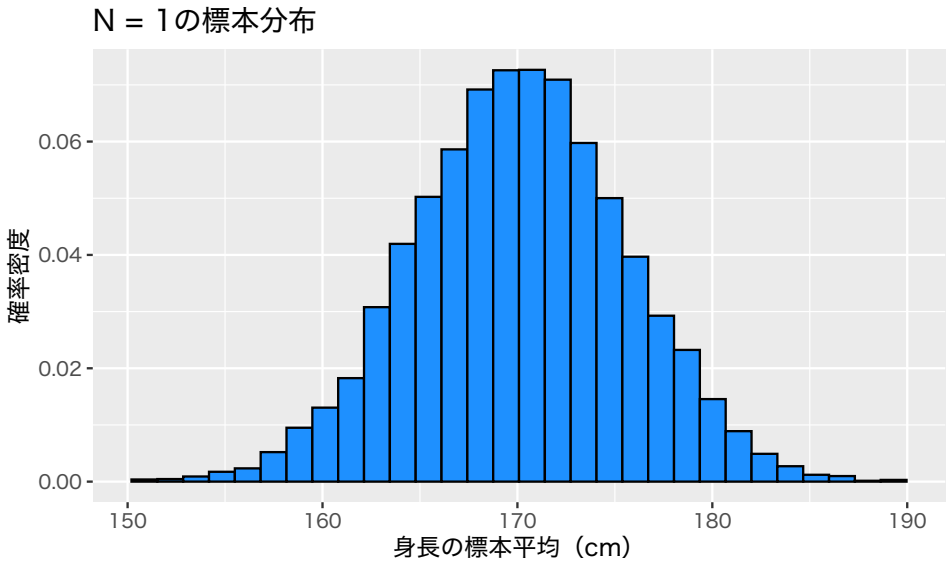
これを 1 万回繰り返す。

```
sim_1 <- rep(NA, n_sims)
for (i in 1 : n_sims) {
  h <- rnorm(N, mean = mu, sd = sigma)
```

```
  sim_1[i] <- mean(h)
}
```

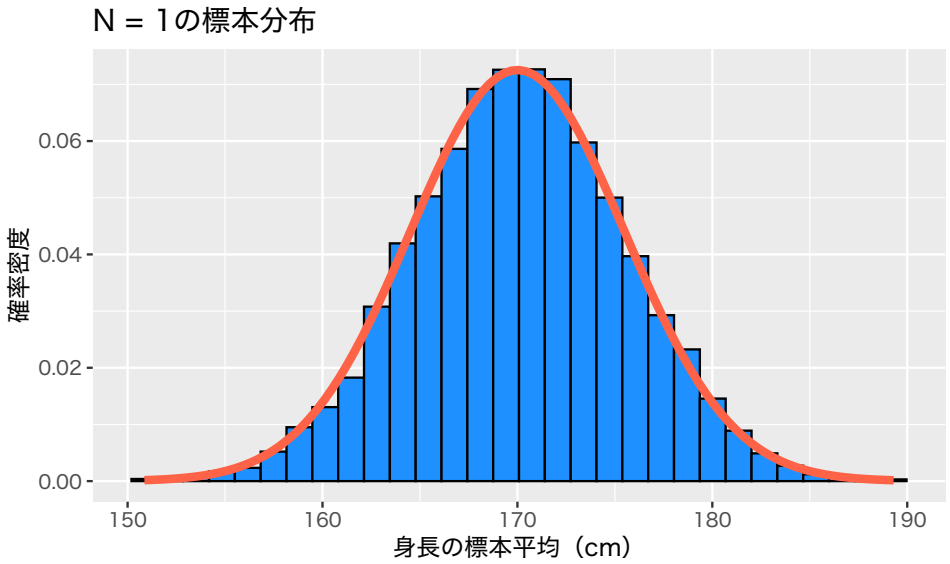
結果をヒストグラムに。

```
df1 <- tibble(sim_1)
hist1 <- ggplot(df1, aes(x = sim_1, y = after_stat(density))) +
  geom_histogram(color = 'black',
                 fill  = 'dodgerblue') +
  labs(x = '身長の標本平均(cm)',
       y = '確率密度',
       title = 'N = 1の標本分布')
plot(hist1)
```




このヒストグラムに、平均 170、標準偏差 5.5 の正規分布（$N(170, 5.5)$）の確率密度曲線を重ねてみる。

```
hist1_2 <- hist1 +
  stat_function(fun         = dnorm,
                args        = list(mean = 170, sd = 5.5),
                inherit.aes = FALSE,
                color       = 'tomato',
                linewidth   = 1.5)
plot(hist1_2)
```




標本平均の標本分布と、母集団の分布はよく似ている。

統計量は、

```
mean(sim_1) # 標本平均の平均値
```

```
[1] 170.1051
```

```
sd(sim_1)   # 標本平均の標準偏差 = 標準誤差
```

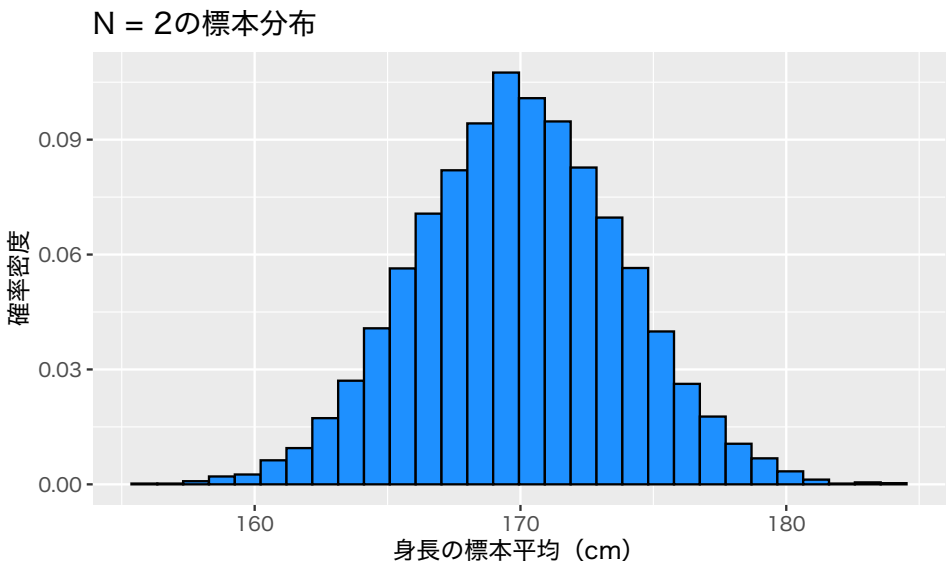
```
[1] 5.453966
```

このように、$N = 1$ の場合には、標本平均の平均値は母平均とほぼ同じ。標本平均の標準偏差（標準誤差）は、母標準偏差とほぼ同じ。

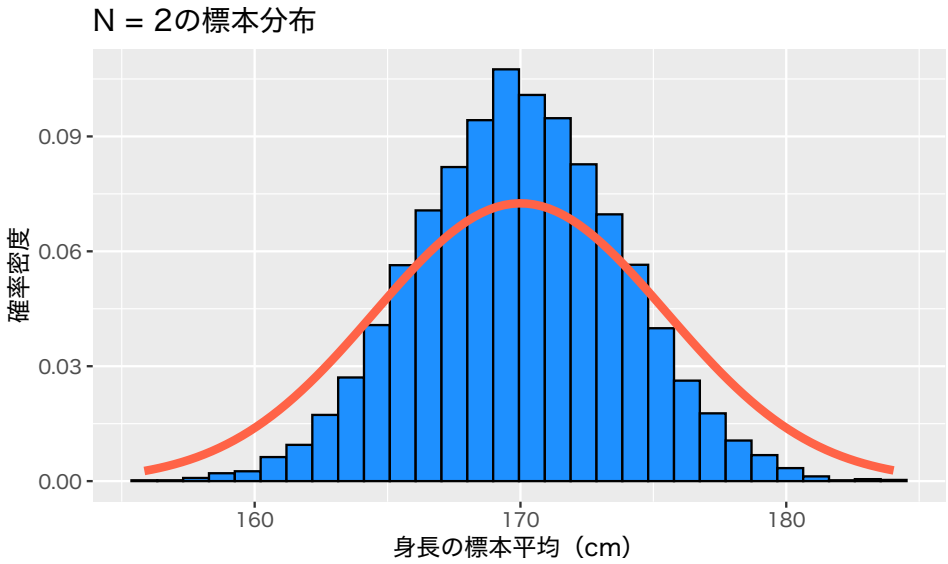
**標本サイズ（$N$）が 2 のとき**

```
N <- 2
sim_2 <- rep(NA, n_sims)
for (i in 1 : n_sims) {
  h <- rnorm(N, mean = mu, sd = sigma)
  sim_2[i] <- mean(h)
}
df2 <- tibble(sim_2)
hist2 <- ggplot(df2, aes(x = sim_2, y = after_stat(density))) +
  geom_histogram(color = 'black',
```

```
                fill  = 'dodgerblue') +
  labs(x = '身長の標本平均(cm)',
       y = '確率密度',
       title = 'N = 2の標本分布')
plot(hist2)
```




```
hist2_2 <- hist2 +
  stat_function(fun         = dnorm,
                args        = list(mean = 170, sd = 5.5),
                inherit.aes = FALSE,
                color       = 'tomato',
                linewidth   = 1.5)
plot(hist2_2)
```




先ほどとは異なり、標本平均の分布と、母集団の分布は少し異なる。標本分布のほうが母集団よりも狭い範囲に集まっていることがわかる。

統計量は、

```
mean(sim_2)
```

```
[1] 169.9444
```

```
sd(sim_2)
```

```
[1] 3.857648
```

標本平均の平均値は母平均とほぼ同じである。しかし、標本平均の標準偏差である標準誤差は、母標準偏差よりもかなり小さくなっている。

理論的には、標本平均の標準偏差（標準誤差）は $\frac{\text{母標準偏差}}{\sqrt{\text{標本サイズ}}}$ になるはず。

```
sigma / sqrt(N)
```

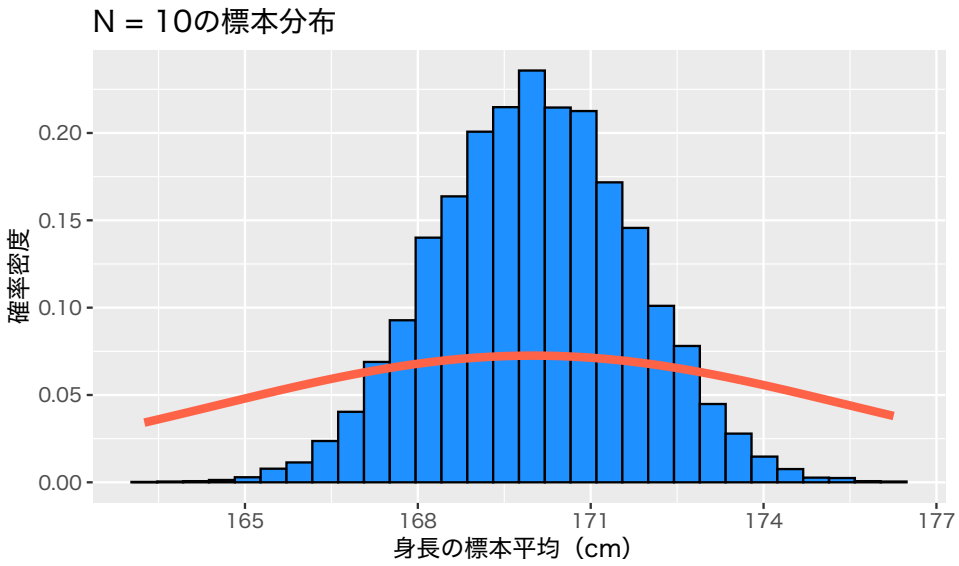
```
[1] 3.889087
```

これは、上で求めた標準誤差とほぼ一致する。

#### 標本サイズ（$N$）が 10 のとき

標本サイズ $N = 10$。

```r
N <- 10
sim_3 <- rep(NA, n_sims)
for (i in 1 : n_sims) {
  h <- rnorm(N, mean = mu, sd = sigma)
  sim_3[i] <- mean(h)
}
df3 <- tibble(sim_3)
hist3 <- ggplot(df3, aes(x = sim_3, y = after_stat(density))) +
  geom_histogram(color = 'black',
                 fill  = 'dodgerblue') +
  labs(x = '身長の標本平均(cm) ',
       y = '確率密度',
       title = 'N = 10 の標本分布')
hist3_2 <- hist3 +
  stat_function(fun         = dnorm,
                args        = list(mean = 170, sd = 5.5),
                inherit.aes = FALSE,
                color       = 'tomato',
                linewidth   = 1.5)
plot(hist3_2)
```




標本分布がさらに狭い範囲に集まっていることがわかる。

```
mean(sim_3)
```

```
[1] 170.0332
```

```
sd(sim_3)
```

```
[1] 1.707583
```

### 誤差の分布

標本平均の誤差は、

$$\text{誤差} = \text{標本平均} - \text{母平均}$$

と表すことができる。また、標本平均の平均値は母平均に等しいので、

$$\text{誤差} = \text{標本平均} - \text{標本平均の平均値}$$

でも同じ。

したがって、上で実行した 3 つのシミュレーションの誤差は、

```
err_1 <- sim_1 - mean(sim_1)
err_2 <- sim_2 - mean(sim_2)
err_3 <- sim_3 - mean(sim_3)
```

である。さらに、それぞれを標準誤差で割る。

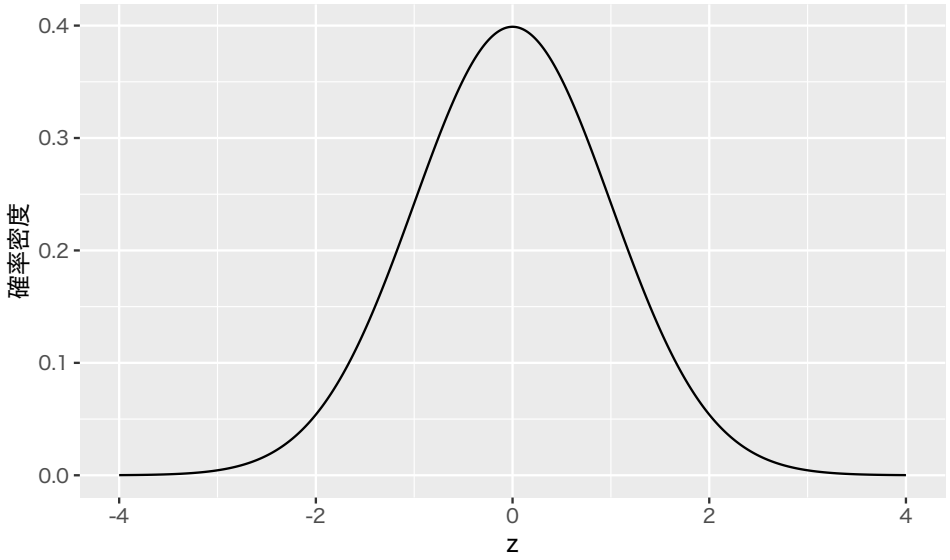
```
z1 <- err_1 / sd(sim_1)
z2 <- err_2 / sd(sim_2)
z3 <- err_3 / sd(sim_3)
```

平均値を引いて、それを標準偏差で割っているので、これは**標準化**（standardization あるいは $z$ 化）である。

ここで、標準正規分布 $\text{Normal}(0, 1)$ の分布を図にしてみる。

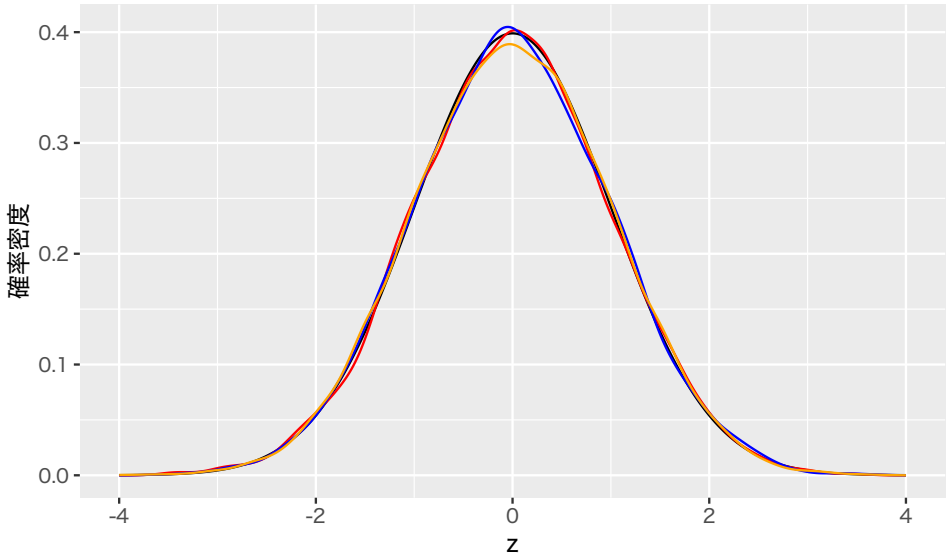
```
df_nml <- tibble(z = seq(from = -4, to = 4, length.out = 1000)) |>
  mutate(dens = dnorm(z, mean = 0, sd = 1))
stdn <- ggplot(df_nml, aes(x = z, y = dens)) +
  geom_line() +
```

```
  labs(y = '確率密度')
plot(stdn)
```




この図に先程計算した z1, z2, z3 の分布を上書きしてみる。確率密度を加えるため、geom_density() を使う。

```
df_z <- tibble(z1, z2, z3)
dens_lines <- stdn +
  geom_density(data  = df_z,
               aes(x = z1, y = after_stat(density)),
               color = 'red') +
  geom_density(data  = df_z,
               aes(x = z2, y = after_stat(density)),
               color = 'blue') +
    geom_density(data  = df_z,
               aes(x = z3, y = after_stat(density)),
               color = 'orange')
plot(dens_lines)
```




このように、標本平均を標準化すると、標準正規分布に似た分布が出てくる。よって、**標準正規分布を利用した推定ができそう**。