# 4. 中心極限定理

honocat

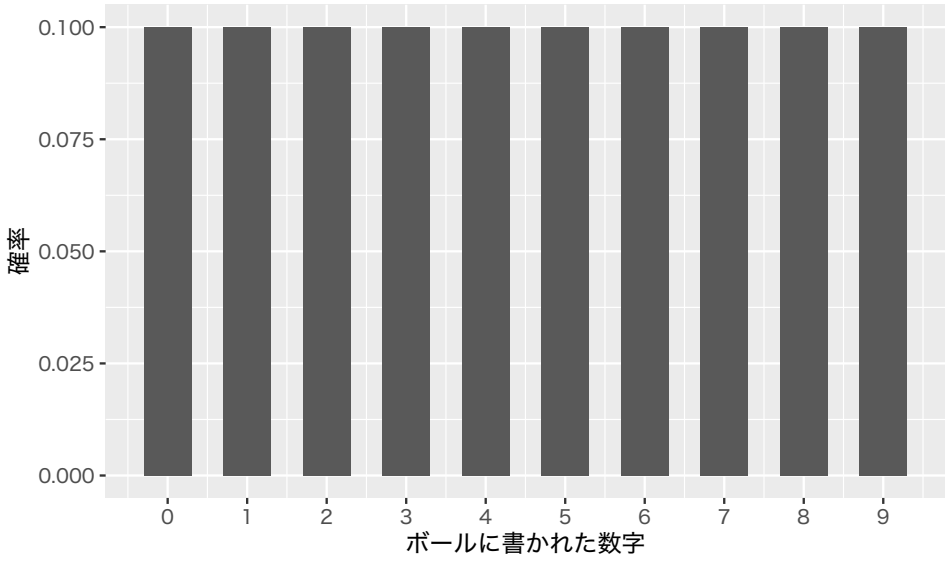
2025-12-16

## 中心極限定理

中心極限定理(Central Limit Theorem)は、「標本サイズ(サンプルサイズ)が大きくなれば、標本平均は正規分布で近似できる」という定理である(より正確には「標本サイズを無限大にすると標本平均を標準化したものが標準正規分布に収束する」という定理である)。

正規分布ではない分布から正規分布ができる。R でシミュレーションを実行することを通じて中心極限定理を理解しよう。

例として、10 個のボールが入った袋を考える。ボールにはそれぞれ 0 から 9 までの数字が書かれているものとする。この袋から、ランダムに 1 つボールを選ぶとすると、選んだボールに書かれた数は、0 から 9 までの整数のどれかで、0 から 9 が選ばれる確率は等しく 10 分の 1 (0.1)である。



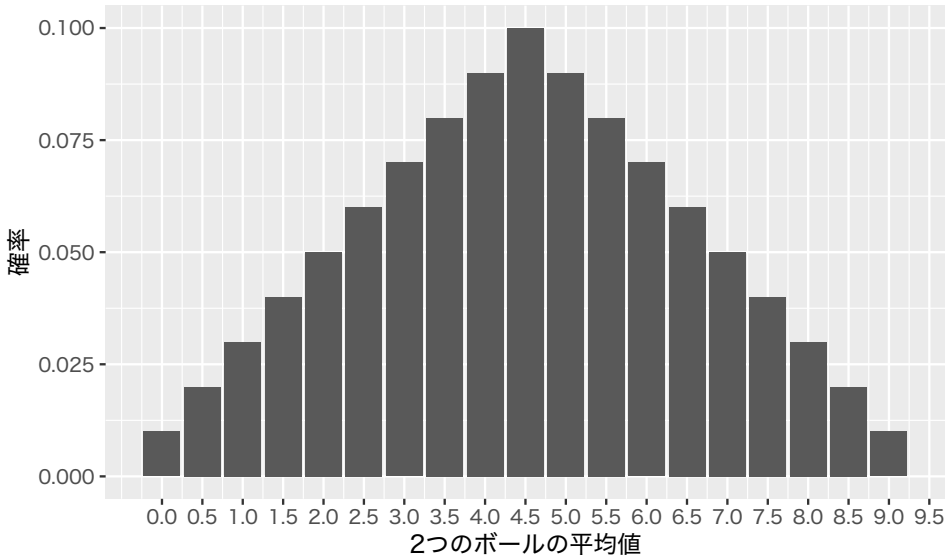

1 つボールを選ぶとき、ボールに書かれている数の平均値(期待値)は、$(9-0)/2 = 4.5$ である。

ここで、私たちはボールにどんな数が書かれているか知らないとする。この状態で、ボールにかかれている数の平均値を当てたい。

もっとも単純な方法は、ボールを $n$ 回引いて、その平均値を当てるという方法である。

まず、ボールを 2 回だけ引いて平均値を当てるという実験をしてみる。この実験をすると、1 回目のボールの選び方は 10 通り、2 回目のボールの選び方も 10 通りあるので、全部で 100 通りの選び方がある。2 つのボールに書かれている数は 0 から 9 までの整数なので、可能な合計値は 0 から 18 までの 19 通りであり、平均値は「合計値/2」なので、平均値も 19 通りしかない。。

理論的には次のような確率で、それぞれの平均値が得られる。




この図から、この実験を 1 回だけ行うとき、正解である 4.5 が選ばれる確率は 0.1 であることがわかる。試しにやってみる。

```
bag <- 0 : 9
exp_2 <- sample(bag, size = 2, replace = TRUE)
mean(exp_2)
```

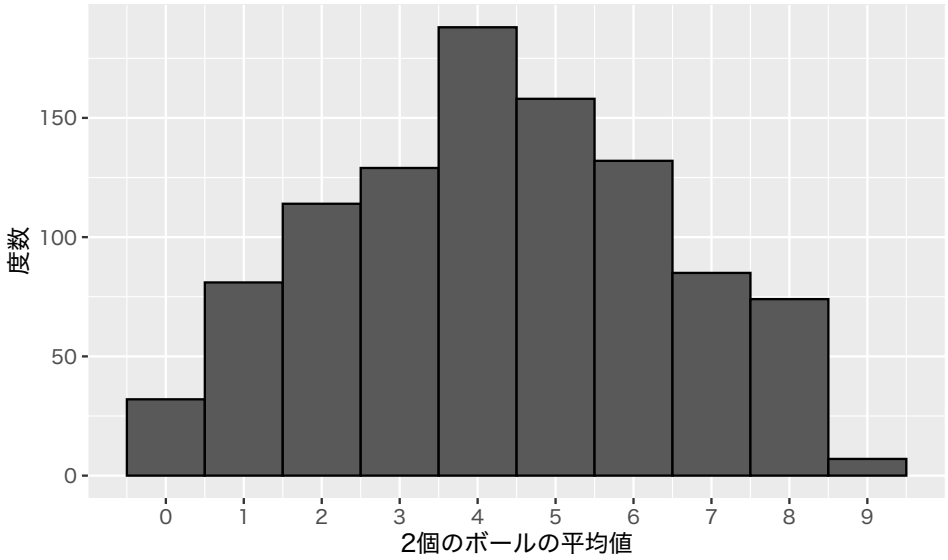
```
[1] 4
```

今回は 4 になった。

では、この実験を 1,000 回繰り返すと「それぞれの回での平均値の分布」はどんな形になるだろうか。

```
N <- 2
trials <- 1000
sim1 <- rep(NA, length.out = trials)
for (i in 1:trials) {
  experiment <- sample(bag, size = N, replace = TRUE)
  sim1[i] <- mean(experiment)
}
```

```
df_sim1 <- tibble(avg = sim1)
h_sim1 <- ggplot(df_sim1, aes(x = avg)) +
  geom_histogram(binwidth = 1,
                 boundary = 0.5,
                 color    = 'black') +
  labs(x = '2 個のボールの平均値',
       y = ' 度数') +
  scale_x_continuous(breaks = 0 : 9)
plot(h_sim1)
```



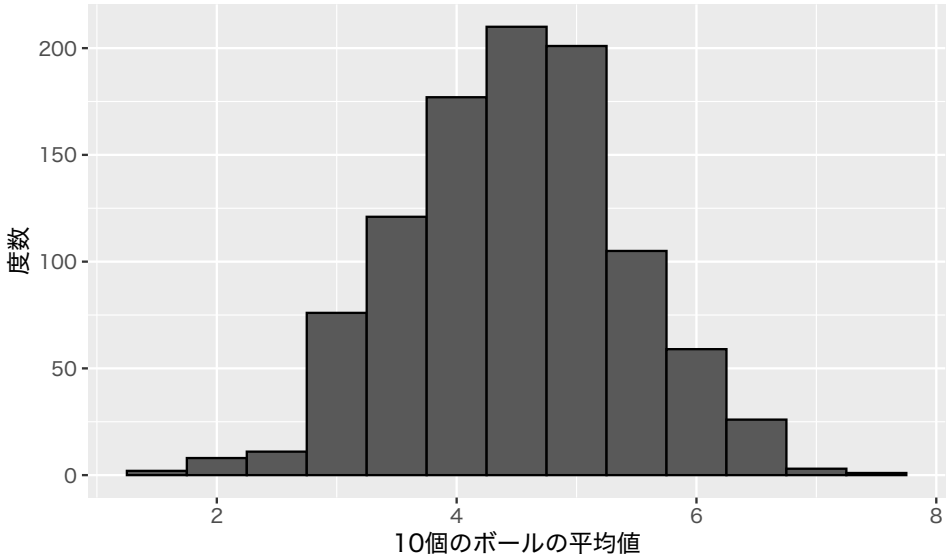

これは正規分布に見えるか？

$N = 10$ でもやってみる。

```
N <- 10
sim2 <- rep(NA, length.out = trials)
for (i in 1:trials) {
  experiment <- sample(bag, size = N, replace = TRUE)
  sim2[i] <- mean(experiment)
}
```

```
df_sim2 <- tibble(avg = sim2)
h_sim2 <- ggplot(df_sim2, aes(x = avg)) +
  geom_histogram(binwidth = 0.5,
                 color    = 'black') +
  labs(x = '10 個のボールの平均値',
```

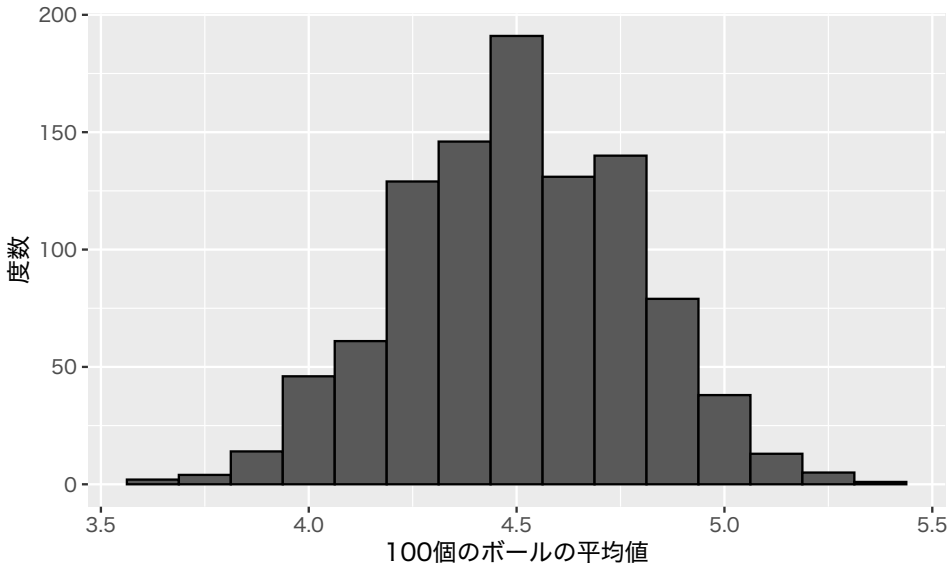
```
       y = '度数')
plot(h_sim2)
```




サンプルサイズを $N = 100$ してみる。

```
N <- 100
sim3 <- rep(NA, length.out = trials)
for (i in 1:trials) {
  experiment <- sample(bag, size = N, replace = TRUE)
  sim3[i] <- mean(experiment)
}
```

```
df_sim3 <- tibble(avg = sim3)
h_sim3 <- ggplot(df_sim3, aes(x = avg)) +
  geom_histogram(binwidth = 0.125,
                 color    = 'black') +
  labs(x = '100 個のボールの平均値', y = '度数')
plot(h_sim3)
```




このように、もとの分布は一様分布でも、サンプルサイズ $N$ を増やすと、「平均値の分布」は正規分布に近づく。よって、サンプルサイズ $N$ が十分に大きい（大まかな目安は $N \geq 100$）とき、正規分布を使って統計的推定や検定を行うことが許されている。