10. 確率分布の代表値 lib

honocat

2025-12-13

離散型分布の例

ベルヌーイ分布

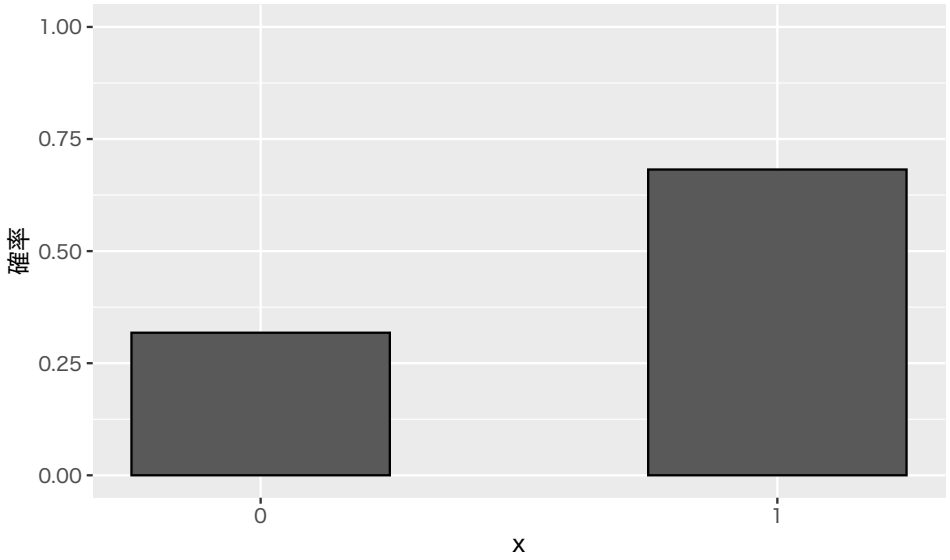
確率変数 X が、

$$X \sim \text{Bernoulli}(\theta)$$

であるとする。

$\theta = 0.7$ の場合について、確率変数 X の値を無作為に 1,000 個生成してみる。

```
x_1 <- rbinom(n = 1000, size = 1, prob = 0.7)
p_x1 <- tibble(x = x_1) |>
  ggplot(aes(x = x, y = after_stat(count) / 1000)) +
  geom_bar(color = 'black',
           width = 0.5) +
  labs(y = '確率') +
  scale_x_continuous(breaks = 0 : 1,
                     minor_break = NULL) +
  ylim(0, 1)
plot(p_x1)
```



x の値は 0 と 1 の 2 種類だけで、$Pr(X = 1) \approx 0.7$ (確率変数 X が 1 となる確率)となっていることがわかる。

x_1 の平均値を計算してみると、

```
mean(x_1)
```

```
[1] 0.682
```

ベルヌーイ分布の分散は $\theta(1-\theta)$ なので、$\theta = 0.7$ の場合は $\mathbb{V}[X] = 0.7(1-0.7) = 0.21$ である。

```
var(x_1)
```

```
[1] 0.2170931
```

近い値である。

二項分布

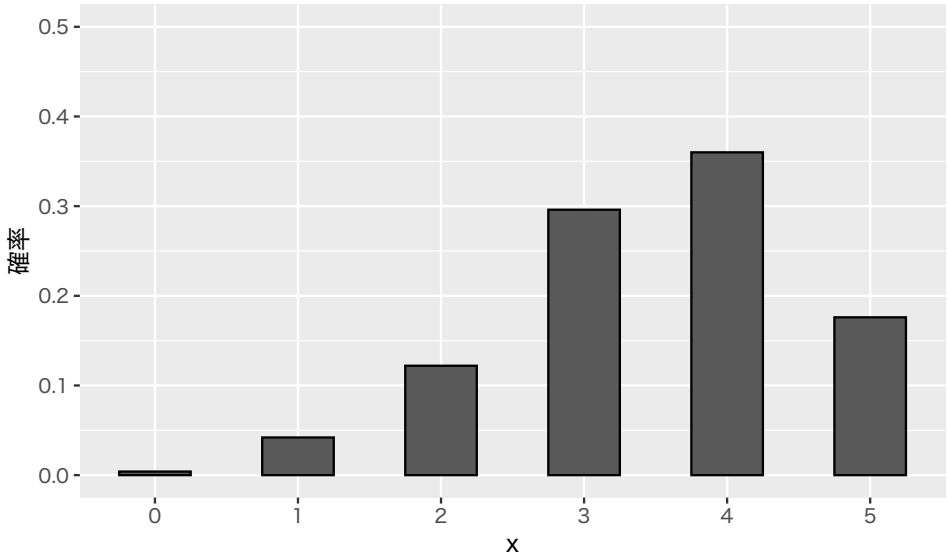
確率変数 X が、

$$X \sim \text{Bernoulli}(n, \theta)$$

だとする。

$n = 5, \theta = 0.7$ の場合について、確率変数 X の値を無策に 1,000 個生成してみる。

```r
x_2 <- rbinom(n = 1000, size = 5, prob = 0.7)
p_x2 <- tibble(x = x_2) |>
  ggplot(aes(x = x, y = after_stat(count) / 1000)) +
  geom_bar(color = 'black',
           width = 0.5) +
  labs(y = '確率') +
  scale_x_continuous(breaks = 0 : 5,
                     minor_breaks = NULL) +
  ylim(0, 0.5)
plot(p_x2)
```



x の値は 0 から 5 までの整数で、$X = 4$ の確率が最も大きく、$X = 0$ の確率が(台の中では)最も小さいことがわかる。

x_2 の平均は、

```r
mean(x_2)
```

```
[1] 3.494
```

期待値である $n\theta = 5 \cdot 0.7 = 0.35$ には一致していないものの、ある程度近い値を取っている。

二項分布の分散は $n\theta(1-\theta) = 5 \cdot 0.7(1-0.7) = 1.05$ である。

```r
var(x_2)
```

```
[1] 1.147111
```

連続型分布の例

正規分布

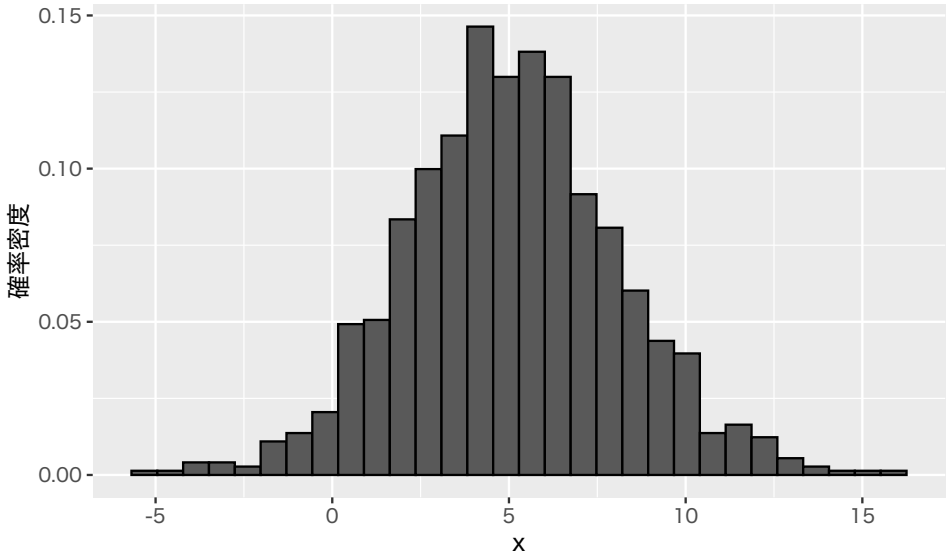
確率変数 X が、

$$X \sim \text{Normal}(\mu, \sigma)$$

であるとする。

$\mu = 5, \sigma = 3$ の場合について、確率変数 X の値を無作為に 1,000 個生成してみる。

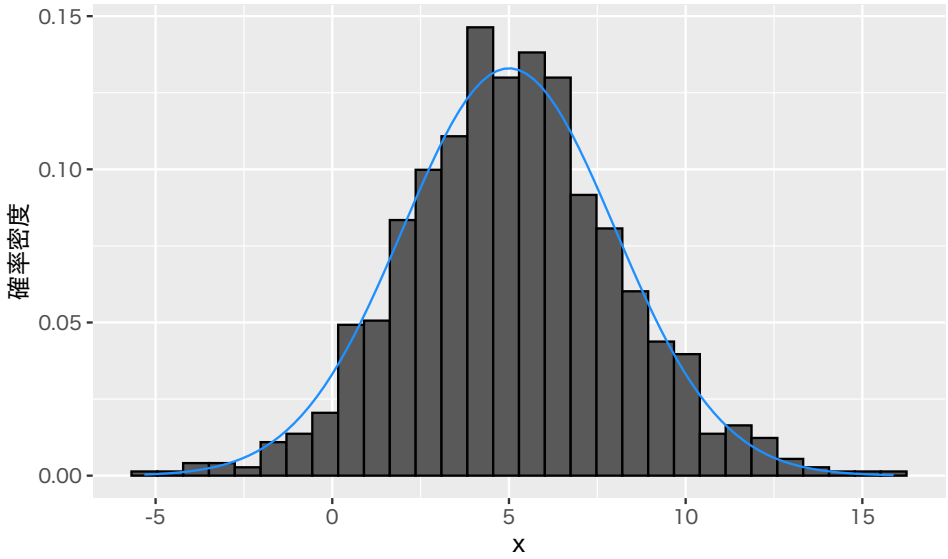
```
x_3 <- rnorm(1000, mean = 5, sd = 3)
p_x3 <- tibble(x = x_3) |>
  ggplot(aes(x = x, y = after_stat(density))) +
  geom_histogram(color = 'black') +
  labs(y = '確率密度')
plot(p_x3)
```



ヒストグラムの概形を見ると、山が一つで、平均値を中心として左右対称になるという正規分布の特徴が見て取れる。

平均が 5、標準偏差が 3 の確率密度曲線を重ねてみる。

```
p_x3_b <- p_x3 +
  stat_function(fun          = dnorm,
                geom         = 'line',
                color        = 'dodgerblue',
                args         = list(mean = 5, sd = 3),
                inherit.aes = FALSE)
plot(p_x3_b)
```



平均値は、

```
mean(x_3)
```

```
[1] 5.079827
```

標準偏差は、

```
sd(x_3)
```

```
[1] 3.032821
```

得られた値のうち、「平均 ±1 標準偏差」「平均 ±2 標準偏差」の範囲に収まっている観測値の数を数える。ただし、得られた値の平均値と標準偏差ではなく、分布の平均と標準偏差を用いる。まず、「平均 ±1 標準偏差」の範囲にある値は、

```
sum(x_3 > 5 - 3 & x_3 < 5 + 3)
```

```
[1] 688
```

である。1,000 個中 688 個の値が「平均 ± 1 標準偏差」お範囲にある。つまり、データの約 68.8% が収まっている。理論的には、この範囲に収まる確率は、

```
pnorm(5 + 3, mean = 5, sd = 3) - pnorm(5 - 3, mean = 5, sd = 3)
```

```
[1] 0.6826895
```

になるはずである。

同様に、「平均 ± 2 標準偏差」の範囲にある個数は、

```
sum(x_3 > 5 - 2 * 3 & x_3 < 5 + 2 * 3)
```

```
[1] 945
```

である。1,000 個中 945 個の値が「平均 ± 2 標準偏差」の範囲にある。つまり、データの約 94.5 が収まっている。この範囲に収まる確率は、

```
pnorm(5 + 2 * 3, mean = 5, sd = 3) - pnorm(5 - 2 * 3, mean = 5, sd = 3)
```

```
[1] 0.9544997
```

になるはずである。

標準正規分布

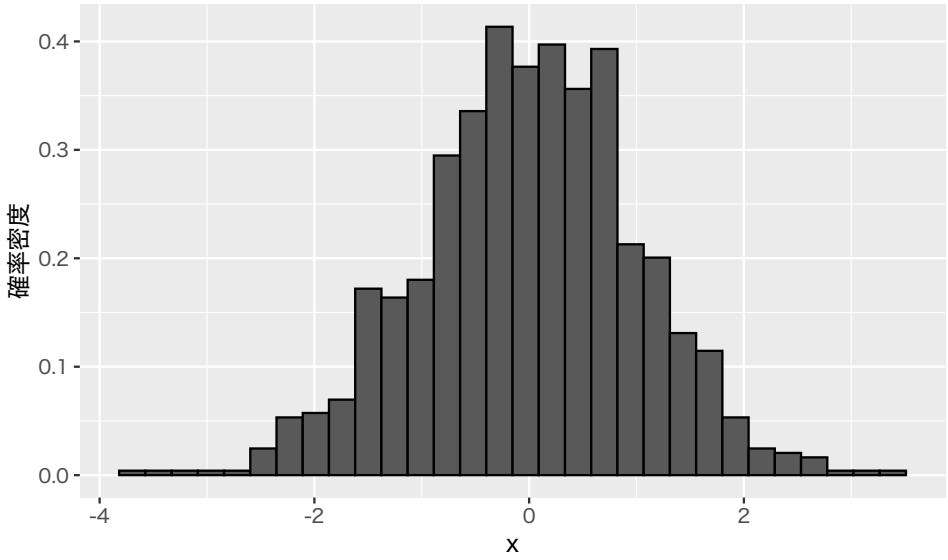
確率変数 X が、

$$X \sim \text{Normal}(0, 1)$$

であるとする。この条件で、確率変数 X の値を無策に 1,000 個生成してみる。

```
x_4 <- rnorm(1000, mean = 0, sd = 1)
p_x4 <- tibble(x = x_4) |>
  ggplot(aes(x = x, y = after_stat(density))) +
  geom_histogram(color = 'black') +
```

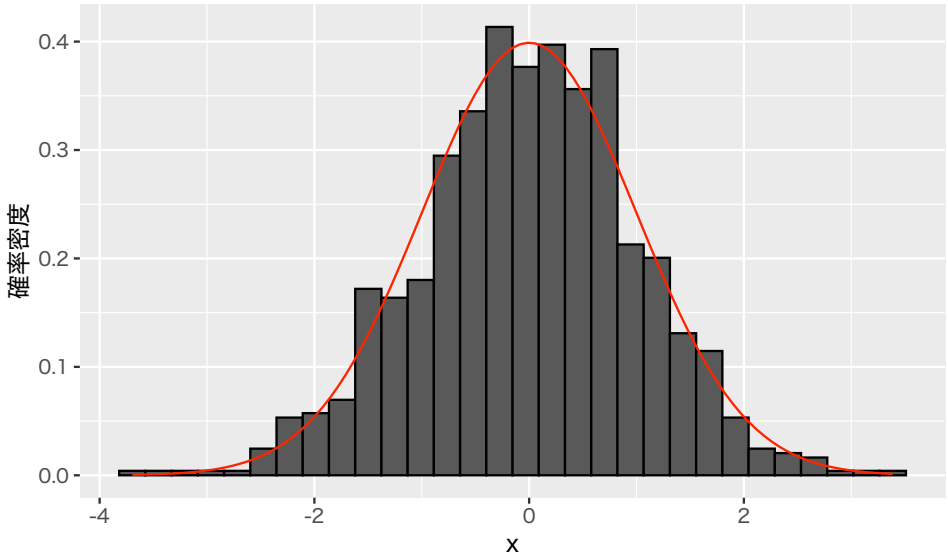
```
  labs(y = '確率密度')
plot(p_x4)
```



ヒストグラムの概形を見ると、山が一つで、平均値を中心として左右対称になるという正規分布の特徴が見て取れる。

上のヒストグラムに標準正規分布の確率密度曲線を重ねてみる。

```
p_x4_b <- p_x4 +
  stat_function(fun          = dnorm,
                geom         = 'line',
                color        = '#FF2400',
                args         = list(mean = 0, sd = 1),
                inherit.aes = FALSE)
plot(p_x4_b)
```



平均値は、

```
mean(x_4)
```

```
[1] -0.01081101
```

標準偏差は、

```
sd(x_4)
```

```
[1] 1.013052
```

得られた値のうち、「平均 ± 1 標準偏差」「平均 ± 2 標準偏差」の範囲に収まっている観測値の数を数えてみる。まず「平均 ± 1 標準偏差」すなわち $-1 < X < 1$ の範囲にある個数は、

```
sum(-1 < x_4 & x_4 < 1)
```

```
[1] 684
```

である。つまりデータの約 68.4% が収まっている。理論的には、この範囲に収まる確率は、

```
pnorm(1, mean = 0, sd = 1) - pnorm(-1, mean = 0, sd = 1)
```

```
[1] 0.6826895
```

になるはずである。

同様に「平均 ± 2 標準偏差」、すなわち $-2 < X < 2$ の範囲にある個数は、

```
sum(-2 < x_4 & x_4 < 2)
```

```
[1] 952
```

である。データの約 95.2% が収まっている。理論的には、

```
pnorm(2, mean = 0, sd = 1) - pnorm(-2, mean = 0, sd = 1)
```

```
[1] 0.9544997
```

になるはず。