9. 確率変数と確率分布

honocat

2025-12-12

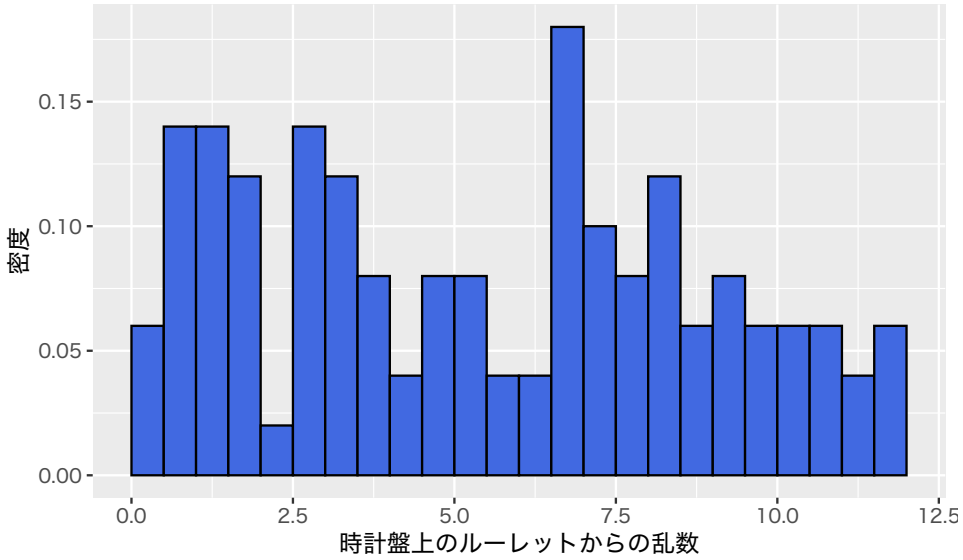
R で確率分布を使う

一様分布

$$X \sim \mathrm{Uniform}(a, b)$$

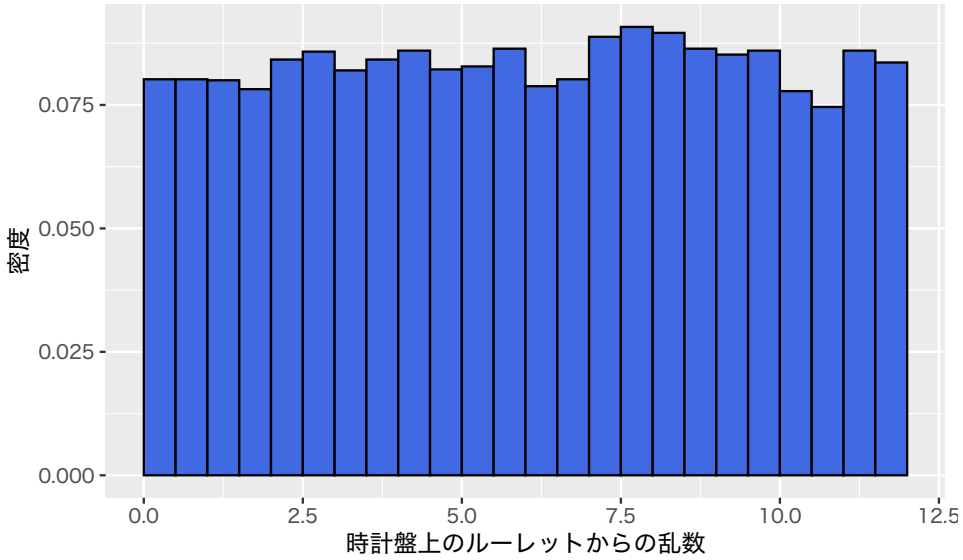
```
a1 <- runif(n = 100, min = 0, max = 12)
```

```
df1 <- tibble(a1)
h1 <- ggplot(df1,
             aes(x = a1,
                 y = after_stat(density))) +
  geom_histogram(binwidth = 0.5,
                 boundary = 0,
                 fill     = 'royalblue',
                 color    = 'black') +
  labs(x = '時計盤上のルーレットからの乱数',
       y = '密度')
plot(h1)
```



生成する乱数の個数を増やしてみる。

```
a2 <- runif(10000, 0, 12)
df2 <- tibble(a2)
h2 <- ggplot(df2,
             aes(x = a2,
                 y = after_stat(density))) +
  geom_histogram(binwidth = 0.5,
                 boundary = 0,
                 fill     = 'royalblue',
                 color    = 'black') +
  labs(x = '時計盤上のルーレットからの乱数',
       y = '密度')
plot(h2)
```



理論的に想定される密度曲線を図に重ねてみる。

```
x_pt <- seq(from = 0, to = 12, length.out = 10000)
```

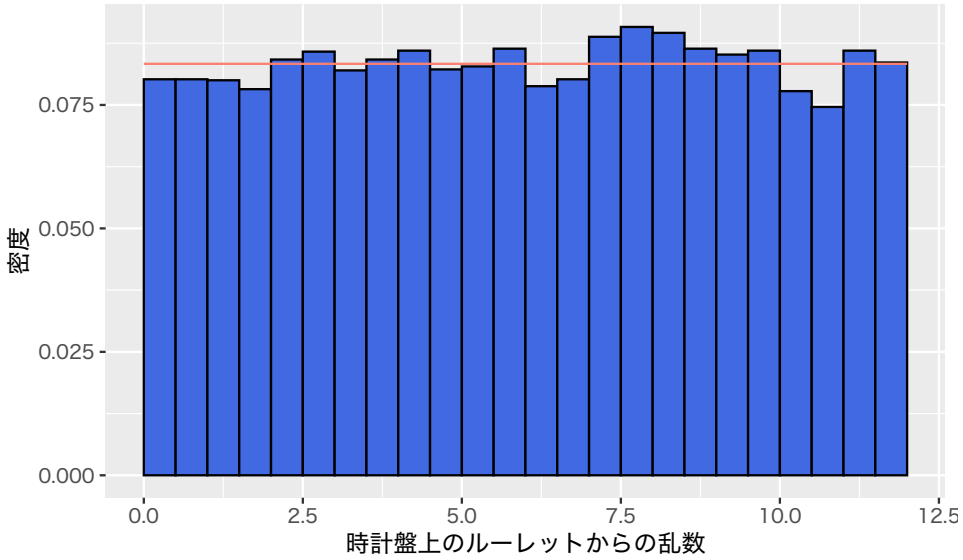
x の各点に対する密度 $f_X(x)$ を求める。

```
dens <- dunif(x_pt, 0, 12)
```

x（x）と dens（$f_X(x)$）をデータフレームに。

```
D_unif <- tibble(x = x_pt, dens = dens)
```

```
h2_b <- h2 +
  geom_line(data = D_unif,
            aes(x = x_pt,
                y = dens),
            color = 'salmon')
plot(h2_b)
```



X がある実数 x 以下の値を取る確率は、分布関数 $F_X(x)$ で求めることができるが、R では次のようにする。$X \leq 3$ となる確率：

```
punif(3, min = 0, max = 12)
```

```
[1] 0.25
```

X が 4.8 以上、11.5 以下になる確率：

```
punif(11.5, 0, 12) - punif(4.8, 0, 12)
```

```
[1] 0.5583333
```

逆分布関数を使うと、分布関数の値が u 以上になる最小の x の値を求めることができる。例えば、分布関数の値が 0.5 以上になるのは、

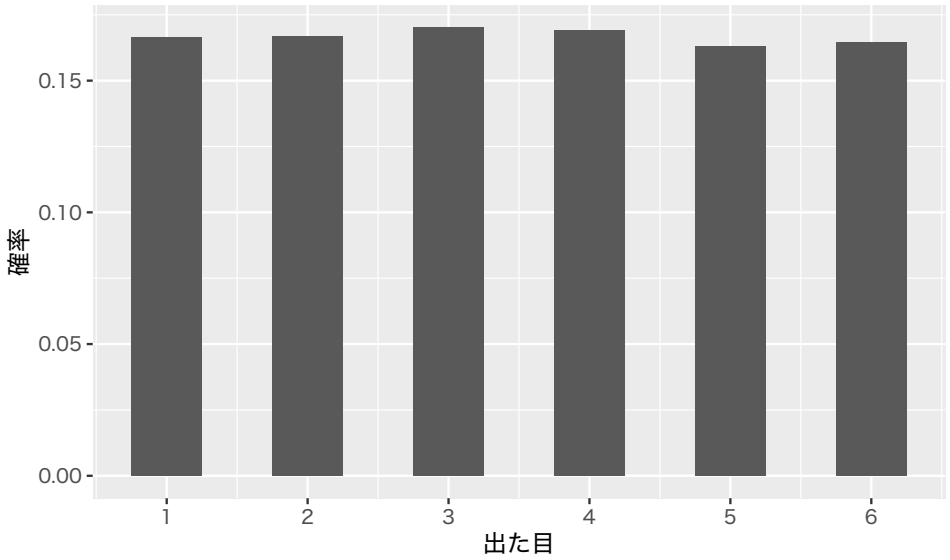
```
qunif(0.5, 0, 12)
```

```
[1] 6
```

である。

連続ではない一様分布（離散一様分布; discrete uniform distribution）からの乱数は、`sample()` で生成できる。

```r
a3 <- sample(1:6, size = 10000, replace = TRUE)
df3 <- tibble(a3)
p3 <- ggplot(df3,
             aes(x = a3,
                 y = after_stat(count) / nrow(df3))) +
  geom_bar(width = 0.5) +
  labs(x = '出た目', y = '確率') +
  scale_x_continuous(breaks = 1:6)
plot(p3)
```



これは棒グラフであり、縦軸は密度ではなく確率。離散型の場合、確率関数(確率質量関数)によって確率変数が特定の値をとる確率を考えることができるので、縦軸に確率をとったグラフを描くことができる。

- 確率関数(確率質量関数) $f(x)$:入力が x だとすると、x を取る確率を返す関数。すべての x における $f(x)$ を足し合わせると 1 になる。
- 確率変数 :数値ではない「事象」を計算可能な「数値」に変換する関数。「1 の目が出た」という結果を「1」という数値に対応させる。

1. 確率変数 X が特定の値をとる

試行が行われる(サイコロを振る)。その結果(事象)を、確率変数 X という関数が数値 x に変換する(例 :「3 の目が出た」という事象→ 3 という数値)。この 3 のように、確率変数 X が取る得る数値を「特定の値」と表現している。

2. 確率を考えることができる

私たちが本当に知りたいのは、その特定の値が出る確率 $P(X = x)$。しかし、確率は事象(結果)に付随するものなので、数値 x から直接確率を計算する関数が必要。

3. 確率関数(確率質量関数)によって

ここで確率関数 $f(x)$。入力値 x が出る確率を返す。

確率関数は「何が起こったか」を数値化し、確率関数はその数値が「どれくらいの頻度で起こるか」を定義する。

二項分布

結果が成功か失敗のみで、成功確率 θ で一定であるような試行を N 回繰り返したとき、その成功確率の分布を「試行回数 N で成功確率 θ の二項分布(binomial distribution)」と呼ぶ。

$$X \sim \text{Binomial}(N, \theta)$$

理論的には、二項分布の平均値(期待値)は $N\theta$、分散は $N\theta(1-\theta)$ になる。したがって、標準偏差は $\sqrt{N\theta(1-\theta)}$ になる。また、最小値は 0、最大値は N。

成功確率 0.4 の試行を 10 回繰り返す実験を 8 回行う。

```
rbinom(n = 8, size = 10, prob = 0.4)
```

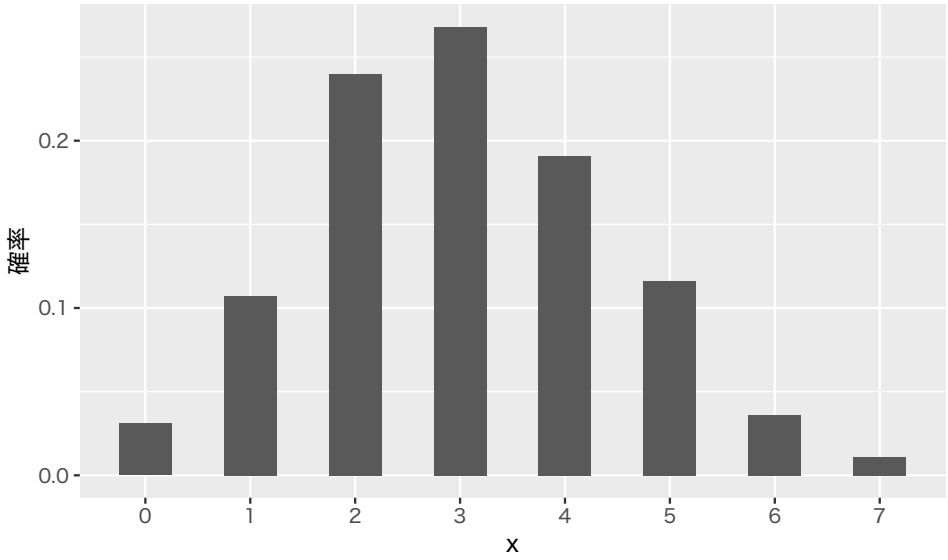
```
[1] 4 4 4 7 7 4 2 5
```

確率変数 X が、$X \sim \text{Binomial}(10, 0.3)$ のとき、確率変数 X をランダムに 1,000 個生成してみる。

```
x <- rbinom(n = 1000, size = 10, prob = 0.3)
```

離散型分布縦軸に確率をとって棒グラフが描ける。

```
D_bin <- tibble(x)
h_binom <- ggplot(D_bin, aes(x = x, y = after_stat(count) / nrow(D_bin))) +
  geom_bar(width = 0.5) +
  scale_x_continuous(breaks = 0 : 10,
                     minor_breaks = NULL) +
  labs(y = '確率')
plot(h_binom)
```



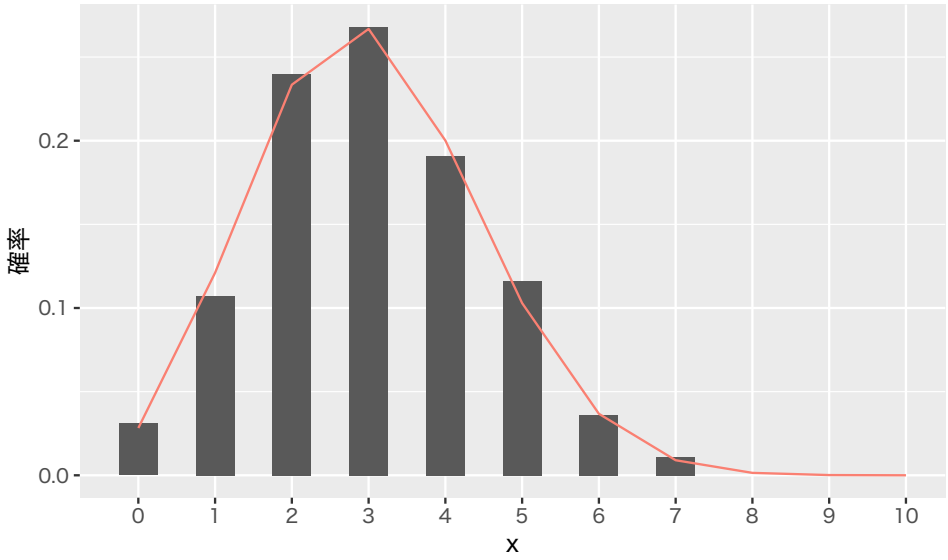
理論的に想定される密度曲線を重ねてみる。X が取りうる値は 0 から 10 まで。

```
x_pt <- 0:10
```

x の各点に対応する密度 $f_X(x)$ を計算する（本当は密度ではなく確率）。

```
dens <- dbinom(x_pt, size = 10, prob = 0.3)
D_binom <- tibble(x = x_pt, dens = dens)

h_binom_b <- h_binom +
  geom_line(data = D_binom,
            aes(x = x, y = dens),
            color = 'salmon')
plot(h_binom_b)
```



X がある実数 x 以下の値を取る確率は、分布関数 $F_X(x)$ で求めることができる。3 以下の確率：

```
pbinom(3, size = 10, prob = 0.3)
```

```
[1] 0.6496107
```

4 以上、6 以下の確率：

```
pbinom(6, size = 10, prob = 0.3) - pbinom(4, size = 10, prob = 0.3)
```

```
[1] 0.1396763
```

逆分布関数。分布関数の値が 0.5 以上：

```
qbinom(0.5, size = 10, prob = 0.3)
```

```
[1] 3
```

正規分布

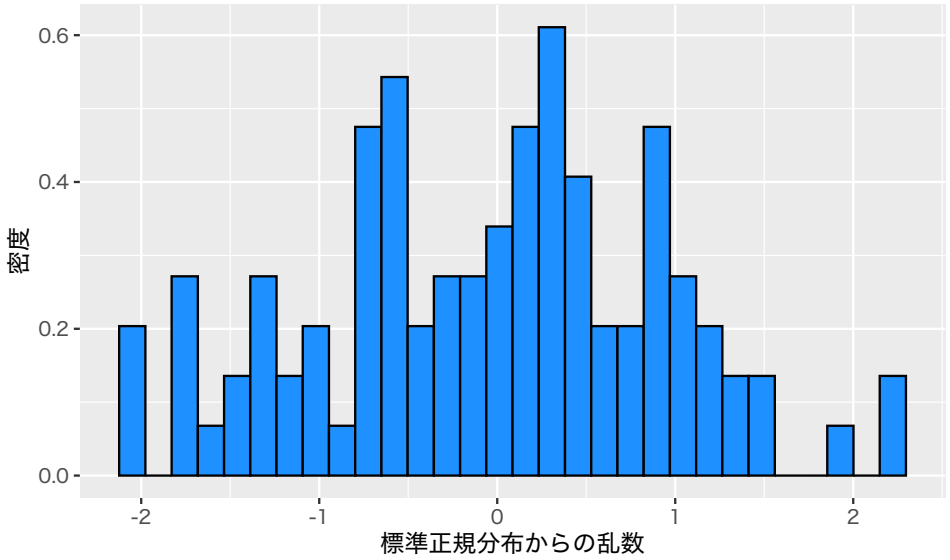
確率変数 X が平均 μ、標準偏差 σ の正規分布に従うとき、

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

標準正規分布から 100 個の乱数を生成してみる。

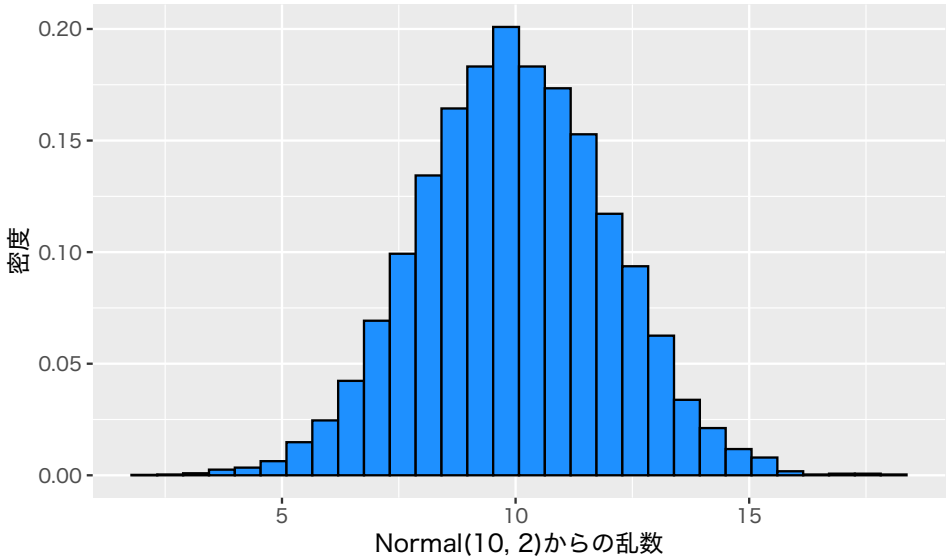
```
b1 <- rnorm(n = 100)

df_n1 <- tibble(b1)
h_n1 <- ggplot(df_n1, aes(x = b1, y = after_stat(density))) +
  geom_histogram(color = 'black', fill = 'dodgerblue') +
  labs(x = '標準正規分布からの乱数', y = '密度')
plot(h_n1)
```



$X \sim \text{Normal}(10, 2)$ とする。

```
b3 <- rnorm(n = 10000, mean = 10, sd = 2)
df_n3 <- tibble(b3)
h_n3 <- ggplot(df_n3, aes(x = b3, y = after_stat(density))) +
  geom_histogram(color = 'black', fill = 'dodgerblue') +
  labs(x = 'Normal(10, 2)からの乱数', y = '密度')
plot(h_n3)
```

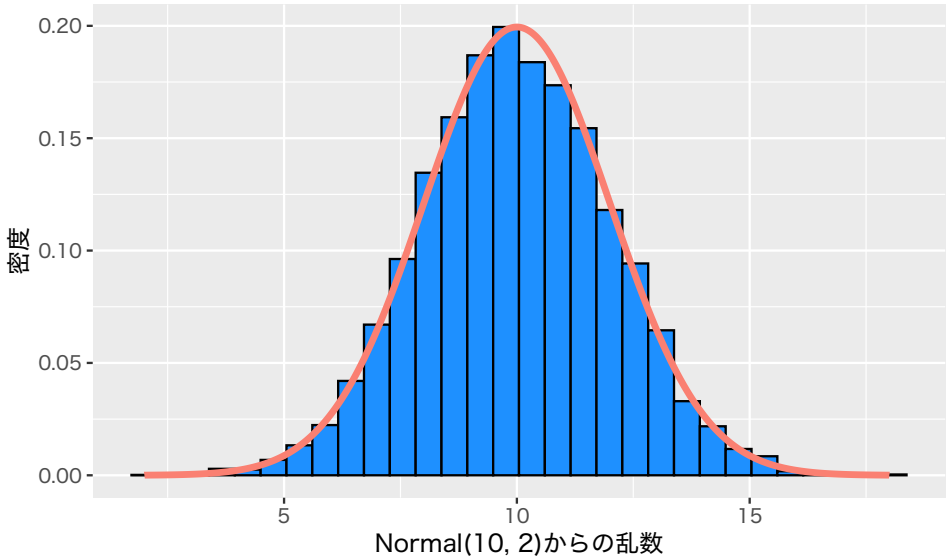


理論的に想定される密度曲線を重ねてみる。

```
x_pt <- seq(from = 2, to = 18, length.out = 10000)
```

x の各点に対応する密度 $f_X(x)$ を計算する。

```
dens <- dnorm(x_pt, mean = 10, sd = 2)
D_nml <- tibble(x = x_pt, dens = dens)
h_n3_b <- h_n3 +
  geom_line(data = D_nml,
            aes(x = x, y = dens),
            color = 'salmon',
            linewidth = 1.2)
plot(h_n3_b)
```



正規分布は実数全体を取り得る。分布関数が 1 となるのは、

```
qnorm(1, mean = 10, sd = 2)
```

```
[1] Inf
```

実数の範囲で $F(x)$ が 1 となることは決してない。