

Rで学ぶ確率統計

第5回 確率分布②

1

Rで学ぶ確率統計



第5回
確率分布②

講師 秋光 淳生

2

第5回 確率分布②

1 多変数の確率分布

2 大標本に関する定理と標本分布

3 Rによるシミュレーション

3

①-1 (復習) 確率分布のまとめ

離散型

- 二項分布 $X \sim B(n, p)$
- 多項分布 $X \sim NB(n, p)$
- 負の二項分布 $X \sim Ge(p)$
- ポアソン分布 $X \sim Po(\lambda)$
- 幾何分布

連続型

- 正規分布 $X \sim N(\mu, \sigma^2)$
- 一様分布 $X \sim U(a, b)$
- ガンマ分布 $X \sim Gam(\lambda, \alpha)$
- 指数分布 $X \sim Exp(\lambda)$

4

①-2 多変数の確率分布(1)

離散型

離散型の変数確率関数

- 同時確率関数
 $P(X = x_i, Y = y_i) = f(x_i, y_i)$
- 周辺確率関数
 $P(X = x_i) = f_x(x_i) \sum_j f(x_i, y_j)$
 $P(Y = y_j) = f_y(y_j) \sum_i f(x_i, y_j)$

多項分布

- k 種類の事象 $A_j: j = 1, 2, \dots, k$
- 各 A_j が確率 p_j で起こる
 X_j : 各 A_j が起こる回数
 $x_j: \{1, 2, \dots\}$
- 同時確率関数
 $P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k)$
 $= \frac{n!}{x_1! x_2! \dots x_k!} p_1^{x_1} p_2^{x_2} \dots p_k^{x_k}$
- パラメータ: n, p_j

Rで学ぶ確率統計

第5回 確率分布②

5

①-2 多変数の確率分布(2)

多変数の確率密度関数

連続型

- 同時確率密度関数

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(u, v) du dv$$

- 分布関数

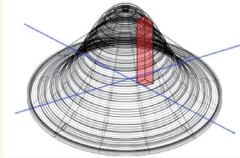
$$P(X \leq x, Y \leq y) = F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv$$

- 互いに独立

$$f(x, y) = f_x(x)f_y(y)$$

$$(-\infty \leq x \leq \infty, -\infty \leq y \leq \infty)$$

2変数の場合
確率は体積を計算



6

①-3 2変量正規分布 (Bi-variate Normal Distribution)

2変量正規分布

- 確率密度関数

$$f(x, y) = \frac{1}{2\pi \sqrt{\sigma_x^2 \sigma_y^2 (1 - \rho^2)}} e^{\{-\frac{1}{2}q(x, y)\}}$$

ただし

$$q(x, y)$$

$$= \frac{1}{1 - \rho^2} \left[\left(\frac{x - \mu_x}{\sigma_x} \right)^2 - 2\rho \left(\frac{x - \mu_x}{\sigma_x} \right) \left(\frac{y - \mu_y}{\sigma_y} \right) + \left(\frac{y - \mu_y}{\sigma_y} \right)^2 \right]$$

- パラメータ: σ_x, σ_y, ρ

- 期待値・分散

$$E[X] = \mu_x$$

$$V[X] = \sigma_x^2$$

$$E[Y] = \mu_y$$

$$V[Y] = \sigma_y^2$$

- 相関係数

$$\text{Cov}[X, Y] = \rho$$

7

①-4 和の分布

和の分布

2つの確率変数 X, Y の和 $X + Y$ について

$$E(X + Y) = E(X) + E(Y)$$

- 期待値

- 分散

$$\begin{aligned} V(X + Y) &= E\{[(X + Y) - (\mu_x + \mu_y)]^2\} \\ &= E[(X - \mu_x)^2] + E[(Y - \mu_y)^2] + 2E[(X - \mu_x)(Y - \mu_y)] \\ &= V[X] + V[Y] + 2\text{Cov}[X, Y] \end{aligned}$$

共分散

- 互いに独立な場合
確率(密度)関数

$$f(x, y) = f_x(x)f_y(y)$$

$$P(X + Y = t) = \sum_k f_x(t - k)f_y(k)$$

$$f(t) = \int_{-\infty}^{\infty} f_x(t - y)f_y(y)dy = \int_{-\infty}^{\infty} f_x(x)f_y(t - x)dx$$

8

①-5 再生性

再生性

- 2つの確率変数 X, Y が独立で、どちらも同じ確率分布に従うときに和 $X + Y$ も同じ確率分布になるとき、再生性を持つという

離散型

$$X \sim B(n_1, p), Y \sim B(n_2, p) \text{ ならば } X + Y \sim B(n_1 + n_2, p)$$

$$X \sim NB(n_1, p), Y \sim NB(n_2, p) \text{ ならば } X + Y \sim NB(n_1 + n_2, p)$$

$$X \sim P_o(\lambda_x), Y \sim P_o(\lambda_y) \text{ ならば } X + Y \sim P_o(\lambda_x + \lambda_y)$$

連続型

$$X \sim \text{Gam}(\lambda, \alpha_x), Y \sim \text{Gam}(\lambda, \alpha_y) \text{ ならば } X + Y \sim \text{Gam}(\lambda, \alpha_x + \alpha_y)$$

$$X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2) \text{ ならば } X + Y \sim N(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$$

Rで学ぶ確率統計

第5回 確率分布②

9

①-6 無記憶性

無記憶性

- ある時刻（ステップ） s (m 回) までに何も起こらなかった条件のもとで、その後の時刻（ステップ） $[s, s+t]$ ($n+m$ 回まで) に起こらない確率

- 幾何分布

$$P(X > n + m | X > m) = (1 - p)^n = P(X > n)$$

- 指数分布

$$P(X > s + t | X > s) = \frac{1 - \int_0^{s+t} \lambda e^{-\lambda x} dx}{1 - \int_0^s \lambda e^{-\lambda x} dx} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t)$$

今まで起こらなかったという条件は、これからの条件に影響を与えない

10

②-1 チェビシェフの不等式と大数の弱法則

チェビシェフの不等式

確率分布に従う確率変数 X が平均が μ 、分散が σ^2 であるとする

このとき

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

が成り立つ

- どんな確率分布でもよい
- $X_1 + X_2 + \dots + X_n$ の分散は $n\sigma^2$
- $\bar{X}$ の標準偏差は $\frac{\sigma}{\sqrt{n}}$
- $k \frac{\sigma}{\sqrt{n}} = \varepsilon$ とすると $\frac{1}{k^2} = \frac{\sigma^2}{n\varepsilon^2}$

大数の弱法則

- 同じ確率分布に従う n 個の確率変数 X_i ($i = 1, 2, \dots, n$) について、それぞれの平均が μ 、分散が σ^2 ($0 < \sigma^2 < \infty$) であるとする

- 標本平均 $\bar{X}$ を

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

とすると任意の ε に対して

$$\lim_{n \rightarrow \infty} P(|\bar{X} - \mu| < \varepsilon) = 1$$

が成り立つ

- これを $\bar{X}$ は母数 μ に確率収束するとい

$$\bar{X} \rightarrow \mu$$

と表す

11

②-2 中心極限定理

中心極限定理

- 同じ確率分布に従う n 個の確率変数 X_i ($i = 1, 2, \dots, n$) それぞれの平均が μ 、分散が σ^2 ($0 < \sigma^2 < \infty$) であるとする
- 標本平均 $\bar{X}$ に対して

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

の分布は $n \rightarrow \infty$ とすると標準正規分布 $N(0, 1)$ に近づく

標本の和に対して

- 個数 n で割ると一定の値に近づく
- 個数の平方根 $\sqrt{n}$ で割ると正規分布に近づく

12

②-3 χ^2 分布 (chi-square Distribution) χ^2 分布

- 標準正規分布に従うデータ

$$Z_1, Z_2, \dots, Z_n$$

に対し

$$W = Z_1^2 + Z_2^2 + \dots + Z_n^2$$

の従う分布を χ^2 分布といい n を自由度という

- 確率変数 X が χ^2 分布に従うことを $X \sim \chi^2(n)$ と書く

- 期待値・分散

$$E[X] = n$$

$$V[X] = 2n$$

- 自由度 m_1 の χ^2 分布に従う確率変数と自由度 m_2 の χ^2 分布に従う確率変数の和は自由度 $m_1 + m_2$ の χ^2 分布に従う
- 同じ正規分布 $N(\mu, \sigma^2)$ に従うデータが n 個あるとき

$$W = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}$$

は自由度 $n - 1$ の χ^2 分布に従う

Rで学ぶ確率統計

第5回 確率分布②

13

②-4 t 分布 (t Distribution)

t 分布

- 確率変数 Z が標準正規分布 $N(0,1)$
- 確率変数 W が自由度 m の χ^2 分布
- 確率変数 Z と W は互いに独立のとき

$$t = \frac{Z}{\sqrt{W/m}}$$

は自由度 m の t 分布に従う

- $X \sim t(m)$ と書く
- 期待値・分散

$$E[t] = 0 \quad (m > 1)$$

$$V[t] = \frac{m}{m-2} \quad (m > 2)$$

- 同じ正規分布 $N(\mu, \sigma^2)$ に従うデータが n 個あるとき

$$Z = \sum_{i=1}^n \frac{(X_i - \mu)}{\sqrt{n}\sigma}$$

$$W = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}$$

とすると (W : 自由度 $n-1$)

$$t = \frac{Z}{\sqrt{W/(n-1)}}$$

は自由度 $n-1$ の t 分布に従う

14

②-5 F 分布 (F Distribution)

F 分布

- 確率変数 W_1 が自由度 m_1 の χ^2 分布
- 確率変数 W_2 が自由度 m_2 の χ^2 分布
- 確率変数 W_1 と W_2 は互いに独立のとき

$$F = \frac{W_1/m_1}{W_2/m_2}$$

は自由度 (m_1, m_2) の F 分布に従う

- $X \sim F(m_1, m_2)$ と書く
- $F(m_1, m_2)$ の上側 (片右側) の $\alpha\%$ の確率を与える点を $F_{\alpha}(m_1, m_2)$ と書く

- それぞれ異なる正規分布に従う

$$X_i \sim N(\mu_1, \sigma_1^2) \quad (i = 1, 2, \dots, n_1)$$

$$Y_j \sim N(\mu_2, \sigma_2^2) \quad (j = 1, 2, \dots, n_2)$$

$$W_1 = \sum_{i=1}^{n_1} \frac{(X_i - \bar{X})^2}{\sigma_1^2}$$

$$W_2 = \sum_{j=1}^{n_2} \frac{(Y_j - \bar{Y})^2}{\sigma_2^2}$$

とすると

$$F = \frac{W_1/(n_1-1)}{W_2/(n_2-1)}$$

は自由度 (n_1-1, n_2-1) の F 分布に従う

15

②-5 F 分布 (F Distribution)

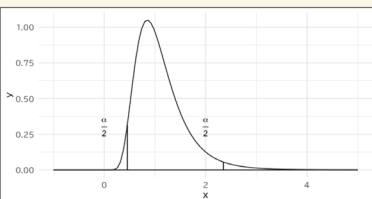
F 分布

- 確率変数 W_1 が自由度 m_1 の χ^2 分布
- 確率変数 W_2 が自由度 m_2 の χ^2 分布
- 確率変数 W_1 と W_2 は互いに独立のとき

$$F = \frac{W_1/m_1}{W_2/m_2}$$

は自由度 (m_1, m_2) の F 分布に従う

- $X \sim F(m_1, m_2)$ と書く
- $F(m_1, m_2)$ の上側 (片右側) の $\alpha\%$ の確率を与える点を $F_{\alpha}(m_1, m_2)$ と書く



とすると

$$F = \frac{W_1/(n_1-1)}{W_2/(n_2-1)}$$

は自由度 (n_1-1, n_2-1) の F 分布に従う

16

③-1 Rにおける分布関数と密度関数の描画

Rにおける関数

dXXXX: 確率密度関数 (確率関数)
 pXXXX: 累積分布関数
 qXXXX: 逆分布関数
 累積分布関数が確率 p になる x
 rXXXX: 乱数
 発生させる個数
 XXXX: 分布の名前

連続型

- χ^2 分布: chisq
(例) dchisq(x, df)
- t 分布: t
(例) dbinom(x, df)
- F 分布: F
(例) dF(x, df1, df2)

```
> ggplot(data=data.frame(x=c(-1,10)), aes(x=x)) +  
  stat_function(fun=df, args=c(df1=30, df2=20))
```

Rで学ぶ確率統計

第5回 確率分布②

17

③-2 大数の弱法則

```
> n <- 50000
> p <- 0.3
> g1 <- tibble(count=1:n,trial= rbinom(n,1,p),
  success=cumsum(trial),prob=success/count )
> ggplot(g1, aes(x=count, y=prob) )+ylim(0,1)+
  geom_hline(yintercept = p, colour="red")+geom_line()
```

- rbinom で1回のベルヌーイ試行
- cumsum 累積和
- prob でその行までの確率
- n回行う

5

18

③-3 中心極限定理

```
n <- 1000
m <- 10000
a <- -10;b <- 10
#
g2 <- numeric(m)
for(i in 1:m){
  g2[i] <- ( mean(runif(n,a,b)) -(a+b)/2 ) / ( sqrt((b-a)^2/12) /sqrt(n))
}
g3 <- tibble(x=g2)
ggplot(g3)+geom_histogram(aes(x=x,y=..density..),bins=50)
```

```
# 一様乱数
平均  $\frac{(a+b)}{2}$  分散  $\frac{(b-a)^2}{12}$ 
# n個の乱数
# m回行う
```

10

15

19

③-4 χ^2 分布

```
n <- 3
m <- 1000
#
g4 <- numeric(m)
for(i in 1:m){
  g4[i] <- ( sum(rnorm(n,0,1)^2) )
}
g5 <- tibble(x=g4)
ggplot(g5)+geom_histogram(aes(x=x,y=..density..),bins=50)+
  stat_function(fun=dchisq, args=c(df=n) )
```

20

25

20

③-5 t 分布

```
a <- 10; b <- 5;
n <- 20; m <- 1000
g6 <- numeric(n)
g7 <- numeric(m)
for(i in 1:m){
  g6 <- rnorm(n,a,b); barx <- mean(g6)
  g7[i] <- sum( (g6-a)/(b*sqrt(n)) )/ sqrt( sum((g6-barx)^2/b^2 ) / (n-1) )
}
g8 <- tibble(x=g7)
ggplot(g8)+geom_histogram(aes(x=x,y=..density..),bins=50)+
  stat_function(fun=dt, args=c(df=(n-1)) )
```

30

35

Rで学ぶ確率統計

第5回 確率分布②

21

3-6 F 分布

```
m <- 1000
n1 <- 60; n2 <- 70;
a1 <- 10; b1 <- 10; a2 <- 30; b2 <- 40
ge <- numeric(m)
for(i in 1:m){
  ga <- rnorm(n1, a1, b1); gb <- rnorm(n2, a2, b2)
  gc <- sum((ga - mean(ga))^2 / (b1^2)); gd <- sum((gb - mean(gb))^2 / (b2^2))
  ge[i] <- (gc / (n1 - 1)) / (gd / (n2 - 1))
}
gf <- tibble(x = ge)
ggplot(gf) + geom_histogram(aes(x = x, y = ..density..), bins = 50) +
  stat_function(fun = df, args = c(df1 = (n1 - 1), df2 = (n2 - 1)))
```